ARI Research Note 93-09

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

Annual Report, 1987 Fiscal Year Supplement to ARI Technical Report 862

Human Resources Research Organization
American Institutes for Research
Personnel Decisions Research Institute
U.S. Army Research Institute

Contracting Officer's Representative
Lawrence M. Hanser

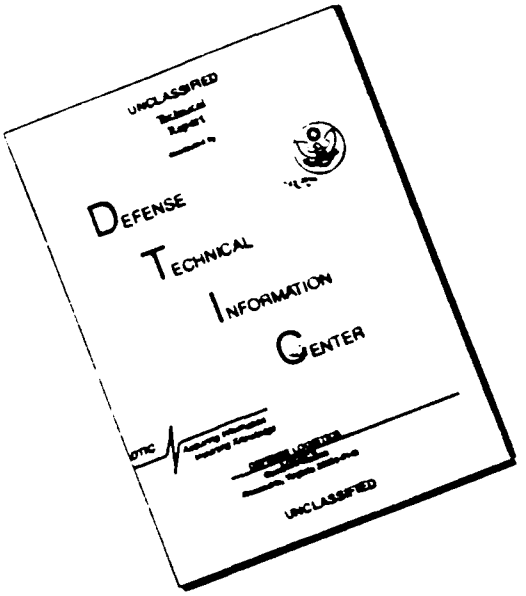
Selection and Classification Technical Area
Michael G. Rumsey, Chief

Manpower and Personnel Research Division
Zita M. Simutis, Director

November 1992

United States Army
Research Institute for the Behavioral and Social Sciences

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
Acting Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Michael G. Rumsey

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1992, November	Interim Oct 86 - Sep 87

4. TITLE AND SUBTITLE Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1987 Fiscal Year Supplement to ARI Technical Report 862

5. FUNDING NUMBERS
MDA903-82-C-0531
63007A
792
232
C71

6. AUTHOR(S) Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, U.S. Army Research Institute for the Behavioral and Social Sciences

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Research Organization
66 Canal Center Plaza, Suite 400
Alexandria, Virginia 22314-4499

8. PERFORMING ORGANIZATION REPORT NUMBER
HumRRO Ir-PRD-88-23

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
Attn: PERI-RS
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARI Research Note 93-09

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

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
--

13. ABSTRACT *(Maximum 200 words)*

The materials presented in this report were prepared under Project A, the U.S. Army's large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. This Research Note supplements the U.S. Army Research Institute for the Behavioral and Social Sciences's Technical Report 862, the project annual report for the 1987 fiscal year. It augments that report by providing copies of a set of technical papers prepared during the year to report on detailed phases of the project research methods and results.

14. SUBJECT TERMS

Army-wide measures	Classification	Criterion measures
Hands-on tests	Knowledge tests	Predictor measures
Project A ratings	Selection	Soldier effectiveness

15. NUMBER OF PAGES 610

16. PRICE CODE --

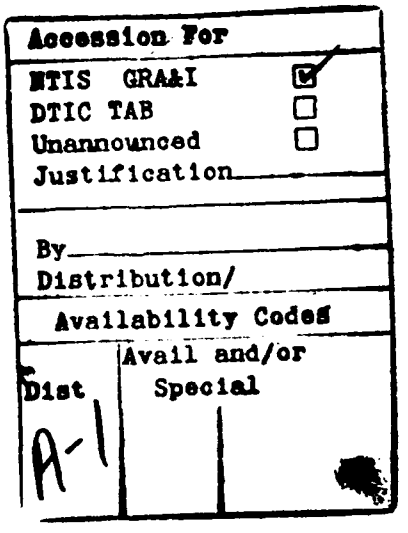
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

EDITORS' PREFACE

In the course of executing the research program of Project A, it has always been an accepted--indeed priority--practice to find mechanisms and means for communicating and sharing early or otherwise salient research results and activities with the U.S. Army and with the professional research community at large. As a result, numerous papers, reports, and symposium proceedings have been produced each year to meet the continuing interest of both scientific and operational audiences. The custom within Project A has been to compile these documents and to publish them as an adjunct to the Project A Annual Report.

The papers in this supplement to the fiscal year 1987 annual report are grouped according to presentation at four professional meetings during the year. Many of the papers are referenced in the annual report. That some are not should in no way diminish their importance or relevance to the readers of these reports. Each document was produced to meet a specific need and audience and, when taken in context, provides in effect a chronology of reports and communications that reveal the process and flow of the overall research program being accomplished collegially by the U.S. Army Research Institute for the Behavioral and Social Sciences and contractor scientists. In many cases these findings have been further refined or synthesized into more formal contract-deliverable items.

Lawrence M. Hanser
Lola M. Zook



IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL

ANNUAL REPORT, 1987 FISCAL YEAR SUPPLEMENT TO ARI TECHNICAL REPORT 862

CONTENTS

CONTENTS (Continued)

CONTENTS (Continued)

IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL

ANNUAL REPORT, 1987 FISCAL YEAR SUPPLEMENT TO ARI TECHNICAL REPORT 862

PURPOSE OF THE REPORT

The materials presented in this report were prepared under Project A, the U.S. Army's current, large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. This Research Note supplements ARI Technical Report 862, the Project Annual Report for the 1987 Fiscal Year. It augments that report by providing copies of a set of technical papers that were prepared during the year reporting on detailed phases of the project research methods and results.

OVERVIEW OF PROJECT A

Project A is a comprehensive long-range research and development program the U.S. Army has undertaken to develop an improved system for selecting and classifying enlisted personnel. The Army's goal is to increase its effectiveness in matching first-tour enlisted manpower requirements with available personnel resources, through use of new and improved selection/classification tests that will validly predict carefully developed measures of job performance. The project addresses the Army's 675,000-person enlisted personnel system encompassing several hundred military occupations.

The program began in 1980, when the U.S. Army Research Institute (ARI) started planning the extensive research needed to develop the desired system. In 1982 ARI selected a consortium, led by Human Resources Research Organization (HumRRO) and including American Institutes for Research (AIR) and Personnel Decisions Research Institute (PDRI), to undertake the 9-year project. It is utilizing the services of 40 to 50 ARI and consortium researchers working collegially in a variety of professional specialties. The Project A objectives are to:

- Validate existing selection measures against both existing and project-developed criteria (including both Army-wide job performance measures based on rating scales, and direct hands-on measures of MOS-specific task performance).
- Develop and validate new selection and classification measures.
- Validate intermediate criteria such as training performance, as predictors of later criteria, such as job performance, so that better informed decisions on reassignment and promotion can be made throughout a soldier's career.

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

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

The research design incorporates three main stages of data collection and analysis in an iterative progression of development, testing, evaluation, and further development of selection/classification instruments (predictors) and measures of job performance (criteria). In the first iteration, file data from fiscal years (FY) 1981/1982 were evaluated to explore relationships between scores of applicants on the Armed Services Vocational Aptitude Battery (ASVAB), and their later performance in training and their scores on first-tour Skill Qualification Tests (SQT).

For the ensuing research, 19 Military Occupational Specialties (MOS) were selected as a representative sample of the Army's 250+ entry-level MOS. The selection was based on an initial clustering of MOS derived from rated similarities of job content. These MOS account for about 45 percent of Army accessions and provide sample sizes large enough so that race and sex fairness can be empirically evaluated in most MOS.

In the second iteration, a Concurrent Validation design was executed with FY83/84 accessions. A "Preliminary Battery" of perceptual, spatial, temperament, interest, and biodata predictor measures was developed and tested with several thousand soldiers as they entered four MOS. The data from this sample were then used to refine the measures, with further exploration of content and format. The revised set of measures was field tested to assess reliabilities, "fakability," practice effects, and other factors. The resulting predictor battery, the "Trial Battery," was administered together with a comprehensive set of job performance indexes based on job knowledge tests, hands-on job samples, and performance rating measures, in the Concurrent Validation during the summer and fall of 1985. The results of the Concurrent Validation were used to form five performance constructs and to report to the Army incremental validities of the Trial Battery components over ASVAB predictors.

On the basis of testing experience, the "Trial Battery" was revised as the "Experimental Predictor Battery," which in turn is being administered in the third iteration, the Longitudinal Validation stage, which began in the late summer of 1986. All measures are being administered in a true predictive validity design. About 50,000 soldiers across 21 MOS are included in the FY86-87 administration and subsequent first-tour measurement. About 3,500 of these soldiers are expected to be available for second-tour performance measurement in FY91. Three MOS were added to the original 19 (19K, 29E, and 96B), and one of the original MOS was dropped (76W).

For administrative purposes, Project A is divided into five research tasks: Task 1, Validity Analyses, and Data Base Management; Task 2,

Developing Predictors of Job Performance; Task 3, Developing Measures of School/Training Success; Task 4, Developing Measures of Army-Wide Performance; Task 5, Developing MOS-Specific Performance Measures.

Activities during the first four years of Project A were reported as follows: FY83, ARI Research Report 1347 and its Technical Appendix, ARI Research Note 83-37; FY84, ARI Research Report 1393 and two related reports, ARI Technical Report 660 and ARI Research Note 85-14; FY85, ARI Technical Report 746 and ARI Research Note (in preparation); FY86, ARI Technical Report 813101 and ARI Research Note 8913704.

Other publications on specific activities during those years are listed in the above reports. The annual report on project-wide activities during FY87 is presented in ARI Technical Report 862. The technical papers reproduced in this Research Note serve as additional documentation for various FY87 activities.

RELATIONSHIP OF SQT SCORES TO PROJECT A MEASURES

Jane M. Arabian and Jeanne K. Mason

U.S. Army Research Institute

Presented on Session, "Test Validation"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Relationship of SQT Scores to Project A Measures

Jane M. Arabian and Jeanne K. Mason
U. S. Army Research Institute
Alexandria, Virginia

The Army develops and administers Skill Qualification Tests (SQT) to soldiers in many of the Military Occupational Specialties (MOS). The testing program was originally intended to diagnose needs for training. However, SQT scores are also used for personnel management decisions (e.g., promotion policy decisions, distribution goals for soldier quality, etc.)

Although SQT are not developed for all MOS, particularly the smaller MOS, the MOS that do have SQT represent a variety of occupational specialties and a large proportion of Army accessions. Further, the test administration and score reporting program is well-established, rendering the SQT scores readily accessible to the Army research community. Since these skill tests are administered to soldiers after school training (AIT), when soldiers have had experience performing in their specialty, the SQT scores have been employed as proxy measures of job performance to support personnel policy decisions. However, the assumption that SQT can be validly used as a measure of job performance has not been tested directly.

Converging evidence does suggest that SQT are viable measures of job performance. For example, the distribution of SQT scores by ASVAB (Armed Services Vocational Aptitude Battery) scores, more specifically Aptitude Area (AA) composite scores from ASVAB, were employed by proponent schools to support particular MOS AA entry score requirements. Along with the proponents' input, the Army's submission to Congress on Army manpower quality goals also included data on the relationship of written and hands-on performance scores, obtained from TRASANA, with ASVAB scores (Office of the Assistant Secretary of Defense, 1985). While both sets of data (SQT/ASVAB and TRASANA data/ASVAB) produced similar results, namely a positive relationship between ASVAB and the performance measures, direct examination of the relationship between SQT and TRASANA hands-on and written test scores was precluded by the small number of cases available with both sets of scores. Consequently, it was not possible to determine the validity of SQT scores as measures of job performance at that time.

With the collection of job performance data from the 1985 (concurrent validation) testing phase of the Army's Project A, "Improving the Selection, Classification and Utilization of Army Enlisted Personnel", and the merging of SQT data into the Project's research database, it has become possible to validate SQT scores against independently developed criteria of job performance. The Project A measures selected for this SQT validation research include paper and pencil measures of school knowledge and job knowledge as well as a work sample (hands-on) measure of job proficiency. If the results of this research demonstrate a strong positive relationship between SQT scores and the Project A measures, then it could be confidently asserted that the SQT are valid measures of job performance. Use of SQT data would then be empirically justified as a measure of job performance for personnel management decisions.

Method

Subjects

The subjects in the present research are a sub-sample of the Project A concurrent validation sample. The data for Project A were collected from June to November 1985. The soldiers were all at Skill Level 1 with 18 to 24 months

experience in the Army at the time of testing. The sub-sample had 3,117 soldiers with test scores for each measure of interest (SQT and three Project A measures as well as ASVAB). Soldiers from the following eight MOS were represented in the sub-sample: 11B--Infantryman; 13B--Cannon Crewman; 19E--Tank Crewman; 31C--Radio Teletype Operator; 63B--Light Wheeled Vehicle/Power Generation Mechanic; 64C--Motor Transport Operator; 71L--Administrative Specialist; 95B--Military Police.

Measures

The SQT is a multiple choice, written test of overall MOS knowledge designed for a 2-hour administration period. Soldiers are tested by MOS and Skill Level. Tasks included in the SQT are randomly selected from the Soldier's Manual for a given MOS. Approximately 20-35 tasks (maximum of 161 items) are included in an SQT. The notice announcing the test includes a list of 150% of the tasks that will appear on the test. The overall SQT score is a percentage computed by adding all scores from each task and dividing the sum by the total number of tasks on the test. Further information about SQT development and administration is available in the SQT Test Development Manual (TRADOC, 1983). SQT scores used in the present research were from the 1985 administration with the exception of MOS 31C, whose scores were from 1986.

The Project A School Knowledge tests (K3), also labelled Job-Relevant Knowledge tests, were developed to measure the cognitive component of training (school) success. Test items were based on, e.g., the Army Occupational Survey Program and Program of Instruction (course curriculum) information for each MOS. All items were reviewed by job incumbents, school trainers and appropriate MOS training proponents for content, accuracy, etc. The K3 test for each MOS contained approximately 150 multiple choice items and was administered in a 2-hour period. A detailed description of the test development procedure and psychometric properties of the tests can be found in R. Davis, G. Davis, Joyner, and de Vera (1985).

The development process and psychometric properties of the Task-Based MOS-Specific, Job Knowledge (K5) and Hands-On (HO), measures are described in C. Campbell, R. Campbell, Rumsey and Edwards (1985). Briefly, the job performance domain for each MOS was determined from several sources, including: the Army Occupational Survey Program results, Soldier's Manual of Common Tasks, MOS-specific Soldier's Manuals, and input from the MOS proponent agency. Subject matter experts provided judgments of task criticality, difficulty and similarity. Separate panels of subject matter experts in each MOS used the judgments to select MOS tasks for K5 measure development. The written K5 measures cover some 30 tasks and have approximately 150-200 multiple choice items which require about 2 hours for administration. The HO measures are a sub-set of 15 of the 30 tasks covered in the K5 measure for each MOS. Aside from logistical constraints (e.g., tasks too hazardous to test), tasks selected for testing in the HO mode entailed physical strength or skilled psychomotor performance, performance within a time limit, many procedural steps, and/or steps that are uncued in their normal sequence.

Data and Analyses

A workfile was created from the Project A longitudinal research database. The workfile contained Skill Level 1 SQT score, average percent correct K3, K5 and HO scores and ASVAB AA composite score for each case (subject). The AA score is used in the Army enlistment process as the primary classification eligibility measure for each MOS. Univariate descriptive statistics and correlation analyses were performed using the SAS statistical package.

Results and Discussion

The univariate descriptive statistics for each performance variable (SQT, K3, K5 and HO) by MOS are presented below. There is satisfactory variance and range in the data to permit further analyses.

Table 1

Descriptive Statistics for Each Variable By MOS

MOS		SQT	K3	K5	HO
11B	M	79.27	80.89	68.16	71.99
	SD	10.75	12.72	11.46	7.73
	N	614	599	594	598
	Min	44	17	25	45
	Max	100	84	96	93
13B	M	73.18	84.83	68.64	63.37
	SD	11.56	11.37	10.69	11.87
	N	847	829	828	808
	Min	28	17	24	24
	Max	100	79	84	91
19E	M	74.30	66.08	62.08	76.08
	SD	8.78	12.73	9.81	8.81
	N	433	418	396	407
	Min	8	20	24	50
	Max	94	86	85	93
31C	M	74.79	80.88	80.68	78.51
	SD	8.72	11.55	10.23	8.68
	N	313	298	288	295
	Min	38	21	27	37
	Max	91	84	82	98
63B	M	62.21	80.54	63.52	84.85
	SD	9.18	12.43	10.76	5.44
	N	525	505	488	472
	Min	23	20	27	62
	Max	86	84	86	96
64C	M	82.17	61.13	80.25	71.69
	SD	7.18	12.27	9.83	8.17
	N	561	547	548	521
	Min	52	20	30	43
	Max	98	84	81	89
71L	M	71.44	80.35	57.06	63.34
	SD	13.58	11.18	10.18	10.16
	N	431	416	421	415
	Min	21	24	30	29
	Max	99	86	84	98
95B	M	78.66	80.79	62.80	78.71
	SD	9.89	10.12	9.56	6.87
	N	628	618	606	603
	Min	49	19	26	48
	Max	95	81	86	85

Correlations were obtained between the appropriate AA composite score, SQT, and each Project A performance measure by MOS for cases with complete data. The correlations, in the table below, are generally consistent with data from other studies. The SQT are positively correlated with the ASVAB AA composite scores as well as with the Project A performance measures. Since the focus of this report is on the relationship between SQT and other measures (i.e., K3, K5 and HO) of job performance, weighted averages using the Fisher z transformation were computed across MOS only for the SQT and Project A correlations and the intercorrelations among the Project A measures. As would be expected, the correlations between same-mode measures (paper and pencil, e.g., SQT:K5, K3:K5) are somewhat higher than the cross-mode (paper and pencil vs hands-on, e.g., K3:HO, SQT:HO) correlations.

Table 2

Correlation Coefficients: Cases With All Variables

MOS	AA COMPOSITE	N	AA:SQT	AA:HO	AA:HS	AA:MP	SQT:HO	SQT:HS	SQT:MP	HO:HS	HO:MP	HS:MP
11B	CO	382	.432	.439	.522	.343	.525	.566	.381	.668	.387	.616
13B	FA	411	.293	.348	.418	.142	.488	.503	.433	.694	.442	.443
19E	CO	315	.508	.494	.577	.389	.565	.616	.395	.726	.341	.691
31C	SC	228	.524	.392	.482	.322	.572	.537	.418	.686	.468	.482
63B	MM	398	.501	.641	.542	.299	.588	.597	.367	.735	.412	.356
64C	MP	467	.490	.435	.473	.323	.391	.465	.374	.634	.368	.444
71L	CL	369	.474	.508	.544	.378	.578	.536	.497	.728	.611	.682
95B	MP	463	.495	.483	.342	.331	.387	.355	.335	.583	.278	.364
	N = 3117											
	r						.583	.517	.395	.676	.489	.479

The correlations between SQT and the three Project A measures were corrected for attenuation and range restriction. The reliability estimates for the Project measures, used for the attenuation correction, are presented below. SQT reliability estimates were not available; therefore, the corrections were based on only the Project A measures.

Table 3

Internal Consistency Reliability Estimates

	Test		
MOS	Hands-on	Job Knowledge	School Knowledge
11B	.54 (682)	.89 (678)	.93 (684)
13B	.75 (612)	.85 (639)	.89 (648)
19E	.63 (474)	.89 (459)	.93 (485)
31C	.79 (341)	.86 (326)	.93 (349)
63B	.52 (569)	.87 (596)	.94 (612)
64C	.64 (648)	.85 (668)	.90 (669)
71L	.73 (494)	.82 (501)	.88 (493)
95B	.58 (665)	.84 (665)	.88 (674)

Note: The second entry (in parentheses) is the sample size.

With respect to the correction for range restriction, a formula was employed which is appropriate for the correlation of a new measure, such as the Project measures, with an existing criterion, the SQT, when selection has been made on a third variable, in this case AA composite score (Guilford, 1965). The correlations between SQT and the Project A measures, corrected for attenuation and range restriction are presented below. Again, weighted averages of the validity coefficients across MOS were computed. It can be seen in the table below that SQT is strongly correlated with each of the independent measures of job performance. The somewhat lower average correlation between SQT and HO scores may be attributable at least in part to measurement mode differences (written vs hands-on).

Table 4

[illegible] With All Variables: Corrected For Attenuation and Range Restriction

MOS	SQT: K3	SQT: K5	SQT: HO
11B	.646	.674	.631
13B	.503	.625	.534
19E	.703	.765	.606
31C	.679	.675	.563
63B	.756	.768	.646
64C	.661	.729	.674
71L	.705	.606	.676
95B	.609	.653	.665
$\bar{X}$	.679	.689	.682

In order to compare scores across the four measures, equi-percentile equating was performed; the results are presented below. Since 60 is used as the passing score for SQT, the percentile for a score of 60 on SQT was used to determine comparable (in terms of percentile) scores for the K3, K5 and HO measures. Thus, 11B soldiers with an SQT score of 60 are in the 6.03 percentile. For the K3 measure, an 11B soldier in the 6.03 percentile would have a score of 37. The percentile for SQT scores of 60, 70 and 80 were determined along with the comparable scores on the Project A measures. Scores for SQT, K3, K5 and HO tests at the 50th and 85th percentile were also calculated.

The lower scores on the Project A measures, compared to the SQT scores, suggest that the Project tests may have been somewhat more difficult. Whether or not the apparent differences in difficulty can be attributed to test content versus the opportunity to study for the test cannot be ascertained. However, it should be noted that SQT test dates with 150% of the tasks to be covered are published before testing; this is not the case with the Project A testing.

Table 5

Equi-Percentile Equating

		Score			
MOS	%ILE	SQT	K3	K5	HO
11B	6.03	60	37	40	50
	16.09	70	40	49	65
	40.86	80	62	61	72
	50	80	63	61	73
	85	90	72	72	80
13B	13.35	60	41	40	50
	40.40	70	52	50	61
	72.50	80	61	60	69
	50	73	56	62	63
	85	86	66	72	76
19E	5.94	60	40	45	62
	27.71	70	61	57	71
	74.13	80	75	69	82
	50	75	60	64	78
	85	82	70	72	85
31C	5.43	60	40	43	56
	32.27	70	56	55	67
	69.05	80	66	65	75
	50	75	61	61	71
	85	83	71	70	79
63B	39.43	60	57	61	84
	81.91	70	70	74	89
	98.86	80	81	83	94
	50	62	60	64	85
	85	71	72	75	90
64C	0.71	60	34	32	47
	5.00	70	39	42	58
	35.12	80	58	56	69
	50	83	63	59	72
	85	90	73	60	79
71L	21.11	60	50	40	58
	43.30	70	58	56	62
	71.23	80	66	63	60
	50	73	60	57	63
	85	86	70	60	73
95B	6.40	60	25	31	40
	8.92	70	43	48	60
	61.94	80	63	66	73
	50	70	60	63	71
	85	84	68	71	77

Note. [illegible] equating [illegible] approximate values rounded to nearest whole numbers.

The equi-percentile equating performed on this data set should not be taken to suggest cut off scores for the Project measures. (Nor would it be reasonable to alter the SQT cut off given only the data presented here). While it would be possible to apply standard setting procedures to the Project A data, it would not be advisable to use the SQT score of 60 to set standards on the other measures. The primary reason for this position is that the SQT cut off score of 60 was not necessarily derived empirically or validated against a definition of minimally acceptable performance. In order to evaluate the SQT cut off, and perhaps determine cut offs on the Project A tests, additional information would be needed about satisfactory and unsatisfactory performance levels.

Conclusions

Project A research has provided a unique opportunity to validate SQT against independently derived measures of job performance. The research presented in this paper strongly supports the validity of SQT as a measure of job performance. Although only a limited number of MOS were in the sample, the variety of occupations and the consistency of the results suggest that SQT in general (i.e., including MOS not in the sample) may serve as a valid measure of job performance for personnel management decisions. Further research is particularly needed, however, to validate the SQT cut off score.

References

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1985). Development and field test of task-based MOS- specific criterion measures (ARI Technical Report 717). Alexandria, VA: Army Research Institute.

Davis, R. H., Davis, G. A., Joyner, J. N., & de Vera, M. V. (1985). Development and field test of job-relevant knowledge tests for selected MOS (ARI Technical Report 757). Alexandria, VA: Army Research Institute.

Gilford, J. P. (1965). Fundamental statistics in psychology and education (Fourth Edition). New York: McGraw-hill Book Company. Office of the Assistant Secretary of Defense (Manpower, Installations and Logistics). (1985). Report to the House and Senate Committee on Armed Services, Quality of Military Enlisted.

SAS Institute, Inc. (1986). Statistical Analysis System, Version 82.4. Cary, NC: SAS Institute, Inc.

TRADOC. (1983). Skill qualification tests (SQT): Policy and Procedures (TRADOC Reg 351-2). Fort Monroe, VA: Department of the Army.

ACKNOWLEDGMENTS

The opinions, views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, expressed or implied, of the U.S. Army Research Institute for the Behavioral and Social Sciences or the Department of Defense or the United States Government. Project A is an Army Research Institute contractual effort, #2Q263731A792, performed by Human Resources Research Organization, American Institutes of Research, and Personnel Decisions Research Institute. The authors wish to express their appreciation to Winnie Young for preparing the data file used in this research.

SKILL REQUIREMENT INFLUENCES ON MEASUREMENT METHOD INTERCORRELATIONS

Charlotte H. Campbell
Human Resources Research Organization

Michael G. Rumsey
U.S. Army Research Institute

Presented on Session, "Issues in Hands-On Performance Testing"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Skill Requirement Influences on Measurement Method Intercorrelations

Charlotte H. Campbell
Human Resources Research Organization

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

The Army is currently engaged in a project, commonly referred to as Project A, to develop a job-based selection and classification system. The project involves the linking of existing and newly developed predictor measures to measures of performance in the Army. The success of the project will depend in no small part on the degree to which the performance measures accurately and comprehensively reflect actual performance of Army jobs. Toward the end of developing a comprehensive performance measurement system, we have developed four different kinds of measures--ratings, administrative measures, hands-on job performance (work sample) measures, and job knowledge measures.

Here we focus on two of the testing methods--hands-on performance tests and job knowledge tests. It has been suggested that, short of measurement in an actual job situation, a hands-on test has the highest fidelity of any type of measure (Vineberg & Taylor, 1978). Yet, probably because of the enormous expense associated with hands-on tests, they are seldom used. Written tests are less costly to administer and in some cases may be as appropriate as, or more appropriate than, hands-on tests. To use an example presented by Vineberg and Taylor (1972), a knowledge test is better suited to assess an automobile driver's knowledge of driving rules and road signs than a hands-on test.

It is of considerable practical interest to know the extent to which the two testing methods are interchangeable. If it could be shown that both methods provide virtually identical information, then one could be eliminated and considerable savings could be achieved. Otherwise, one must consider the possibility that each type of measure provides a unique, valid contribution to an overall assessment of an incumbent's job proficiency and that both are needed to obtain maximum job coverage.

An investigation by Rumsey, Osborn and Ford (1985) used meta-analytic procedured to examine the relationship between hands-on and job knowledge tests. Excluding investigations which used a language-oriented work sample, they found a mean correlation of .57, adjusted for attenuation, between hands-on and job knowledge tests. This correlation suggests some degree of overlap but not total interchangeability.

Are there factors which might substantially moderate the correlation between the two types of measures? Rumsey, et al. (1985) found some evidence

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

that type of work sample had an impact, as correlations for investigations using verbal performance tests tended to exceed those in investigations using motor performance tests. These investigators also found limited support for the proposition that type of occupation influences the correlation obtained. However, much remains to be learned about potential moderating factors.

Vineberg and Taylor (1972) have suggested that the extent to which a job requires skill is an important consideration in examining correlations between knowledge and work sample measures. They noted that skill, unlike knowledge, can only be acquired through practice. Job knowledge tests are presumably best suited to measure knowledge; performance tests are presumably best suited to measure job skills. For those jobs in which task requirements can be reduced to job knowledge, the correspondence between the two types of measures should be high; for those in which skill is an important requirement, the correspondence should be lower.

The effort reported here involved first identifying the skills that are required to perform hands-on tasks that are tested in nine military occupational specialties (MOS) in Project A. Then, the extent to which these requirements moderate correlations between job knowledge and hands-on test scores was determined.

Method

Occupations (MOS). Performance tests and job knowledge tests were developed for nine Army occupations, or Military Occupational Specialties (MOS). These MOS were selected to be as representative of the full set of entry-level MOS as possible, covering the range of job content, Career Management Fields, and ASVAB Aptitude Area prerequisites. The MOS are shown in Table 1.

Task Selection. For each MOS, selection of tasks from the job domain proceeded according to four criteria: the tasks should cover the job content areas, they should be the relatively more important ones, they should permit variability of performance, and they should not be of very low performance frequency.

Test Development. Fifteen tasks in each MOS were selected for performance testing based on such factors as number of cued steps and degree of skill required. Performance tests were developed to score the soldier on whether each step of the task was performed correctly, and to provide standard conditions and instructions for the testing. Multiple-choice format job knowledge tests were also developed for those tasks in each MOS. All tests were pilot-tested, and later field-tested on 114 to 178 soldiers in each MOS. Results from those administrations were used to revise the tests; in some cases, hands-on tests or job knowledge tests were dropped.

Data Collection. Between June and November, 1985, the hands-on and knowledge tests were administered to over 5000 skill level 1 soldiers in the nine MOS, at 14 sites in the U.S. and Europe. (This was Project A's Concurrent Validation phase.) The numbers of soldiers tested in each MOS are shown in Table 1. Job knowledge tests were administered by project staff; actual scoring of the performance tests was done by NCO, trained in scoring procedures by project staff.

Table 1

MOS Selected for Testing and Numbers of Soldiers Tested

MOS		Number Tested
11B	Infantryman	662
13B	Cannon Crewman	586
19E	Tank Crewman	434
31C	Single Channel Radio Operator	303
63B	Light Wheel Vehicle Mechanic	541
64C	Motor Transport Operator	629
71L	Administrative Specialist	481
91A	Medical Specialist	480
95B	Military Police	638

Knowledge/Proficiency Assignments. Three project staff who had been involved in test development and had served as hands-on test managers during the Concurrent Validation testing independently sorted the hands-on steps into one of the three categories: knowledge, simple motor, or complex motor. The level of agreement among the judges was around 80% across the nine MOS; disagreements were resolved by discussing the assignments among the three judges.

Because each performance test score was the percent of steps performed correctly, we classified the tests as K (Knowledge) if at least half of the steps had been sorted into the knowledge category, and as P (Proficiency) if half or more of the steps were in the two proficiency categories. The P tasks were further categorized as P1 (simple motor tasks where manipulation is trivial, easy to perform, and easily learned) if more steps were in the P1 category than in either of the other two categories, or as P2 (complex motor tasks which require more than two trials to perform well) if more steps were in the P2 category than either of the other two categories. Tasks where the number of P1 and P2 steps were the same, or where neither P1 nor P2 outnumbered the K steps, were held out of analyses that compared those two levels of categorization.

Table 2 shows the number of tasks in each MOS that were tested in both the performance mode and the job knowledge mode, and the number of tasks where the performance test was categorized as K, P1, or P2.

Data Analysis. The nine MOS had between 14 and 17 tasks tested in both the job knowledge and performance modes. For each task, the scores used were the percent of steps performed correctly and the percent of items answered correctly. These scores were then correlated by task across the soldiers in each MOS. After the correlations were transformed to Fisher z scores, they were entered into an analysis of variance, with the nine MOS and the knowledge/proficiency categories as independent variables.

Results

Table 3 presents the means and standard deviations of the correlations between performance tests and job knowledge tests for each of the nine MOS;

Table 2

Number of Tasks Tested in Performance and Job Knowledge Modes and Number of Tasks Assigned to Knowledge/Proficiency Categories for Nine MOS

MOS		Tasks	K	P1	P2	Total P[a]
11B	Infantryman	12	2	7	2	10
13B	Cannon Crewman	17	2	8	7	15
19E	Tank Crewman	14	5	7	1	9
31C	Single Channel Radio Operator	15	10	4	0	5
63B	Light Wheel Vehicle Mechanic	15	4	4	6	11
64C	Motor Transport Operator	14	3	5	5	11
71L	Administrative Specialist	12	4	1	7	8
91A	Medical Specialist	15	6	6	1	9
95B	Military Police	16	8	4	3	8

[a]Includes tasks not clearly P1 or P2; see text.

the statistics are also shown for the groupings of tasks based on knowledge/proficiency category assignments. (The correlations had been transformed, using the Fisher z transformation, before calculating the summary statistics; the results shown in Table 3, however, have been transformed back to Pearson correlations.) In eight of the MOS, the individual task correlations ranged from about .00 to .40; in one MOS, the highest correlation was .19. (Task correlations tend to be substantially lower than correlations for entire jobs; hence, the level of these correlations cannot be meaningfully compared with earlier findings.) With the large number of soldiers tested in each MOS, even small correlations (around .08) are significant at the .05 level. Over two-thirds of the correlations in every MOS were significant at that level.

Two analyses of variance were calculated, using the transformed correlations (Fisher z) as the dependent variable. In the first ANOVA, the nine MOS and the two knowledge/proficiency categories (K and P) were the independent variables. The second ANOVA likewise used MOS, and also the three levels of the knowledge/proficiency categorization (with two levels of proficiency - simple motor (P1) and complex motor skills (P2), as the independent variables. Both ANOVA results are summarized in Table 4.

In both analyses, the main effect for MOS was nonsignificant, and the interaction terms were not significant. In both analyses, the knowledge/proficiency term was significant. Where knowledge/proficiency was considered on only two levels, the difference favored the K tasks, where the performance test had been categorized as predominantly knowledge. In the second analysis, where there were three groups of tasks - knowledge (K), simple motor (P1), and complex motor (P2) - comparisons of the means of those groups revealed that only the difference between K tasks and P1 tasks was significant at the .01 level ($F = 14.33$, $df = 2,95$); K tasks and P2 tasks differed slightly ($F = 6.68$, $df = 2,95$, $p < .10$), as did K tasks and the combined group of P1 tasks and P2 tasks ($F = 7.581$, $df = 3,95$, $p < .10$). The difference between P1 and P2 tasks was not one bit significant.

Table 3

Means and Standard Deviations of Performance x Job Knowledge Test Correlations by Knowledge/Proficiency Category for Nine MOS

MOS			Tasks	K	P1	P2	Total P
11B	Infantryman	N	12	2	7	2	10
		Mean	.17	.26	.18	.09	.15
		S.D.	.37	.14	.16	.02	.14
13B	Cannon Crewman	N	17	2	8	7	15
		Mean	.17	.20	.16	.17	.16
		S.D.	.11	.07	.11	.13	.12
19E	Tank Crewman	N	14	5	7	1	9
		Mean	.14	.23	.09	.12	.10
		S.D.	.13	.19	.07	-	.06
31C	Single Channel Radio Operator	N	15	10	4	0	5
		Mean	.20	.22	.15	-	.15
		S.D.	.14	.17	.03	-	.03
63B	Light Wheel Vehicle Mechanic	N	15	4	4	6	11
		Mean	.10	.10	.07	.10	.10
		S.D.	.04	.04	.02	.03	.04
64C	Motor Transport Operator	N	14	3	5	5	11
		Mean	.15	.26	.11	.09	.12
		S.D.	.12	.20	.09	.05	.09
71L	Administrative Specialist	N	12	4	1	7	8
		Mean	.24	.30	.16	.20	.20
		S.D.	.11	.13	-	.09	.09
91A	Medical Specialist	N	15	6	6	1	9
		Mean	.17	.17	.15	.33	.17
		S.D.	.13	.18	.08	-	.09
95B	Military Police	N	16	8	4	3	8
		Mean	.15	.18	.10	.10	.11
		S.D.	.11	.11	.06	.17	.10
	Across MOS	N	130	44	46	32	86
		Mean	.16	.21	.13	.14	.14
		S.D.	.12	.15	.10	.11	.10

Table 4

Analysis of Variance Summary Tables for MOS x Knowledge/Proficiency

MOS x Knowledge/Proficiency

	SOURCE	SS	df	MS	F-Ratio	F	p
[1]	MOS	.156	8	.020	[1/4]	1.45	<.25
[2]	K/P	.137	1	.137	[2/3]	18.08	<.01
[3]	MOS x K/P	.061	8	.008	[3/4]	.57	NS
[4]	Within cell	1.499	112	.013			

MOS x Knowledge/Simple Motor/Complex Motor

	SOURCE	SS	df	MS	F-Ratio	F	p
[1]	MOS	.142	8	.018	[1/4]	1.20	NS
[2]	K/P1/P2	.108	2	.054	[2/3]	4.90	<.05
[3]	MOS x K/P1/P2	.161	15[a]	.011	[3/4]	.73	NS
[4]	Within cell	1.388	95	.015			

[a]Reduced by 1 df for missing cell estimation.

Discussion

There is fairly clear evidence here that the differentiation between knowledge requirements and proficiency requirements on hands-on performance tests explains some of the variability in correlations between the two modes of testing. When the steps required on the performance tests are primarily knowledge mediated, and are demonstrations of the acquisition of task knowledge, then the correlations with written tests of the tasks are higher than when most of the performance test steps require demonstration of psychomotor skill, however simple.

Further analyses, already underway, will involve meta-analysis of the obtained correlations, and an examination of the knowledge/proficiency distinction as a possible moderator variable.

REFERENCES

Rumsey, M. G., Osborn, W. C., & Ford, P. (1985). Comparing work sample and job knowledge measures. Paper presented at the annual conference of The American Psychological Association, Los Angeles.

Vineberg, R. & Taylor, E. N. (1972). Performance in four army jobs by men at different aptitude (AFQT) levels: 4. Relationships between performance criteria (Technical Report 72-23, pp. 17, 19). Alexandria, VA: Human Resources Research Organization.

Vineberg, R., & Taylor, E. N., (1978). Alternatives to performance testing: Tests of task knowledge and ratings (Professional Paper 6-78). Alexandria, VA: Human Resources Research Organization.

THE DEVELOPMENT OF A MODEL OF THE PROJECT A CRITERION SPACE

John P. Campbell
Human Resources Research Organization

Lawrence M. Hanser
U.S. Army Research Institute

Lauress Wise
American Institutes for Research

Presented on Symposium,

"Project A Concurrent Validation: Preliminary Results"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

THE DEVELOPMENT OF A MODEL OF THE PROJECT A CRITERION SPACE[1]

John P. Campbell
University of Minnesota

Lawrence M. Hanser
Army Research Institute

Lauress Wise
American Institutes for Research

Conceptual Background

The goals of performance measurement in Project A are to define, or model, the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor. The performance measures are to serve as criteria for validating selection/classification tests, and not, at this point, as operational appraisals.

Some additional specific goals are to: a) make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency, b) compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e., a multi-trait, multi-method approach), c) develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance, and d) evaluate existing archival and administrative records as possible indicators of job performance.

Given these intentions, the criterion development effort focused on three major methods: hands-on job sample tests, multiple choice knowledge tests, and ratings. The behaviorally anchored rating scale (BARS) procedure was extensively used in the development of the rating methods.

Modeling Performance

The development efforts to be described were guided by a particular "theory" of performance. The basic outline is as follows.

First, job performance really is multi-dimensional. There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization (Army).

Two General Factors

For the population of entry level enlisted positions we postulated that there are two major types of job performance components. The first are specific to a particular job. That is, measures of such components would reflect specific technical competence or specific job behaviors that are not

[1]This research was funded by the U. S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U. S. Army Research Institute or the Department of the Army.

required for other jobs. We anticipated that there would be a relatively small number of distinguishable factors of technical performance that would be a function of different abilities or skills.

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

Factors vs. a Composite

Saying that performance is multi-dimensional does not preclude using just one index of an individual's contributions to make a specific personnel decision (e.g., select/not select, promote/not promote). As argued by Schmidt and Kaplan (1971) some years ago, it seems quite reasonable for the organization to scale the importance of each major performance factor relative to a particular personnel decision that must be made and to combine the weighted factor scores into a composite that represents the total contribution or utility of an individual's performance, within the context of that decision.

A Structural Model

If performance is characterized in the above manner, then a more formal way to model performance is to think in terms of its latent structure, postulate what that might be, and then resort to a confirmatory analysis. Within limits, this is what we tried to do. Unfortunately, it is true that we simply know a lot more about predictor constructs than we do about job performance constructs. There are volumes of research on the former, and almost none on the latter.

Unit vs. Individual Performance

Finally, people do not usually work alone. Individuals are members of work groups or units and it is the unit's performance that frequently is the most central concern. Project A has not incorporated unit effectiveness in its model of performance. The project is focused on the development of a new selection/classification system for entry level personnel and is concerned with improving personnel decisions about individuals and not units. The task is to maximize the average payoff per individual selected.

What we have chosen to do is to try to identify the factors, or means, by which individuals contribute to unit performance and to assess individual performance on those factors via rating methods.

Criterion Development

Actual criterion development proceeded from two basic types of information. First, all available task descriptions were used to generate a population of job tasks for each MOS. The principal sources of task

description are the Army's periodic job description surveys and the Soldier's Manual for each MOS which is a specification by management of what the task content of the job is supposed to be. After much editing, revising to insure non redundancy and a uniform level of generality, and a formal review by a panel of subject matter experts, a population of 130-180 tasks was enumerated for each MOS.

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

Standardized job samples, the paper-and-pencil job knowledge tests, and numerical ratings scales were then constructed to assess knowledge and proficiency on these tasks. Each measure went through multiple rounds of pilot testing and revision. The job sample tests were fairly elaborate and were composed of multiple test stations sometimes spread over a football field size area. Because of time limitations (4 hours), only 15 of the tasks could be tested hands-on.

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

Since one major objective was to determine the relationships between training performance and job performance and their differential predictability, if any, a comprehensive training achievement test was constructed for each MOS by carefully matching the content of the program of instruction (POI) with the content of the population of job tasks, and writing items to represent each segment of the match.

The final entry in the array of criterion measures was produced by a concerted effort to get what we could from the files or archival records. We began by enumerating all possibilities from three major sources of such records: the enlisted master file, the enlisted military personnel file, and the military personnel records jacket (the 201 File).

We systematically compared these three sources using a sample of 750 people and a standardized information recording form. The 201 file looked the most promising in terms of recency and completeness, but of course, it is by far the most expensive to search. As a consequence, we collected eight archival performance indicators via a self report questionnaire. That is, people were asked what was in their personnel file as regards letters of commendation, disciplinary actions, etc. Field tests on a sample of 500 people showed considerable agreement between self report and archival records, for both positive and negative things. Further follow-up questionnaires and interviews suggested that self report may be the more accurate. The self report items were combined into five indicators that were actually used as criterion measures.

Determining Actual Criterion Scores

The first step in our analyses was to identify the basic criterion scores whose structure we would analyze. If all the rating scales are used separately and the MOS-specific measures are aggregated at the task or instructional module level, there are approximately 200 criterion scores on each individual. Some aggregation was needed.

Reduction of the Hands-On and Written Variables

The 30 tasks sampled for each job were clustered via expert judgment into 8 to 15 functional categories on the basis of similarity of task content. Each of the school knowledge items was similarly mapped into a specific functional category.

Ten of the functional categories were common to some or all of the jobs (e.g., first aid, basic weapons, field techniques). Each job also had two to five functional performance categories that were unique.

After category scores were computed, separate factor analyses were executed for each type of measure within each job. There were several common features in the results. First, the unique functional categories for each job tended to load on different factors than the common functional categories. Second, the factors that emerged from the common functional categories tended to be fairly similar across the nine different jobs and across the three methods.

Using the empirical factor analysis to guide us, we adopted a set of content categories which became the performance test scores used in subseuqent analyses.

Reduction of the Rating Variables

The individual rating scales were, for the most part, highly reliable. Empirical factor analyses of the Army-wide rating scales suggested three factors. These were:

1. Effort/Leadership, including effort and competence in performing job tasks, leadership, and self-development.
2. Maintaining Personal Discipline, including self-control, integrity, and following regulations.

3. Physical Fitness and Military Bearing, including physical fitness and maintaining proper military bearing and appearance.

Similar factor analyses were reviewed for the job-specific scales for each job. Two factors were identified based on these results. The first consisted of those aspects of job performance that were central to the specific technical content of each job. The second factor included the remaining, less central job performance components.

The individual items in the combat performance prediction battery also were subjected to an empirical factor analysis. Two factors emerged. The first factor consisted of items depicting exemplary effort, skill, or courage under stressful conditions. The second factor consisted of negatively worded items portraying failure to follow instructions and lack of discipline under stressful conditions.

Building the Target Model

The next step was to build a target model of job performance that could be tested for goodness of fit within each of our nine jobs. The project began with an initial model of performance (Borman, Motowidlo, Rose, & Hanser, in press) which had been modified on the basis of field test data (Campbell & Harris, 1985). Principal components factor analyses within MOS were used to suggest further modifications.

Several consistent results were observed. First, the expected "method" factors appeared, specifically one factor for the ratings and one for the written tests. The evidence for a "hands-on" method factor was less compelling. Second, the nature of the substantive factors tended to be similar across MOS.

Based on the empirical analyses, a revised model was constructed to account for the correlations among our performance measures. This model included five job performance constructs and two measurement method factors.

Confirming the Model Within Each Job

The next step in the analysis was to conduct separate tests of goodness of fit of this target model within each of the nine jobs. This was done using the LISREL confirmatory factor analysis program (Joreskog & Sorbom, 1981).

In conducting a confirmatory factor analysis with LISREL, it is necessary to specify the structure of three different parameters matrices: the hypothesized factor structure matrix (a matrix of regression coefficients for predicting the observed variables from the underlying latent constructs); the matrix of uniqueness of error components (and intercorrelations); and a matrix of covariance among the factors. In these analyses, we set the diagonal elements of the covariance matrix to one, forcing a "standardized" solution. This meant that the off-diagonal elements would represent the correlations among and between our performance constructs and method factors. We further specified that the correlation among the two method factors and each performance construct should be zero. This effectively defined the method factor as that portion of the common variance among measures from the same method that was not predictable from (i.e., correlated with) any of the other related factor or performance construct scores.

To be perfectly clear, the approach we used was obviously not purely confirmatory. the hgypothesized target model was based in part on analyses of these same data.

Confirmation of the Overall Model

Given the certain amount of prior examination of the data described above, the results of the confirmatory procedures applied to each job seemed to support a common structure of job performance. The procedures also yielded reasonably similar estimates of the intercorrelations among the constructs an of the loading s of the observed variables on these constructs across the none jobs.

The final step in our analyses was to determine whether the variation in some of these parameters across jobs could be attributed to sampling variation. The specific model that we explored stated that: (1) the correlation among factors was invariant across jobs and (2) the loadings of all of the Army-wide measures on the performance constructs and on the rating method factor were also constant across jobs.

The overall model fit extremely well. The root mean square residual was .047, and the chi-square was 2508.1. There were 2403 degrees of freedom after adjusting for missing variables and the use of the data in estimating uniqueness. This yields a significance level of .07, not enough to reject the model.

Summary and Discussion

Some aspects of the final structure are noteworthy. First, in spite of some confounding with measurement method, the latent performance structure appears to be composed of very distinct components. It is reasonable to expect that the different performance constructs should be weighted in forming an overall appraisal of performance for use in personnel decisions. Using regression techniques to partial the methods factors from the substantive factors should also tell us more about what does or does not predict the residual variance.

Finally, since (a) the five-factor solution is stable across jobs sampled from this population, (b) the performance constructs seem to make sense, and (c) the constructs are based on measures carefully developed to be content valid, it seems safe to ascribe some degree of construct validity to them.

References

Borman, W. C., Motowidlo, S.J., Rose, S. R., & Hanser, L. M. (in press). Development of a model of soldier effectiveness (Technical Report 741). Alexandria, VA: U.S. Army Research Institute.

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

Joreskog, K. C., & Sorbom, D. (1981). LISREL VI: Analysis of Linear Squares methods. Uppsala, Sweden: University of Uppsala.

Schmidt, F. L., & Kaplan, L. B. (1977). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology, 24, 419-434.

Table 1

Summary of Criterion Measures Used in Batch A and Batch Z Concurrent Validation Samples

Performance Measures Common to Batch A and Batch Z

- Army-Wide Rating Scales (all obtained from both supervisors and peers).
  - Ten behaviorally anchored rating scales (BARS) designed to measure factors of non-job-specific performance.
  - Single scale rating of overall effectiveness.
  - Single scale rating of NCO potential.
- Combat prediction scale containing 41 items.
- Paper-and-Pencil Test of Training Achievement developed for each of the 19 MOS (130-210 items each).
- Personnel File Information form developed to gather objective archival records data (awards and letters, rifle marksmanship scores, physical training scores, etc.).

Performance Measures for Batch A Only

- Job Sample (Hands-On) tests of MOS-specific task proficiency.
  - Individual is tested on each of 15 major job tasks in an MOS.
- Paper-and-pencil job knowledge tests designed to measure task-specific job knowledge.
  - Individual is scored on 150 to 200 multiple-choice items representing 30 major job tasks. Ten to 15 of the tasks were also measured hands-on.
- Rating scale measures of specific task performance on the 15 tasks also measured with the knowledge tests. Most of the rated tasks were also included in the hands-on measures.
- MOS-specific behaviorally anchored rating scales (BARS). From 6 to 10 BARS were developed for each MOS to represent the major factors that constitute job-specific technical and task proficiency.

Performance Measures for Batch Z Only

- Army-Wide Rating Scales (all obtained from both supervisors and peers).
  - Ratings of performance on 11 common tasks (e.g., basic first aid).
  - Single scale rating on performance of specific job duties.

Auxiliary Measures Included in Criterion Battery

- A Job History Questionnaire which asks for information about frequency and recency of performance of the MOS-specific tasks.
- Work Questionnaire - a 44-item questionnaire scored on 14 dimensions descriptive of the job environment.
- Measurement Method Rating obtained from all participants at the end of the final testing session.

Table 2

Six basic functional categories of job performance and knowledge obtained from factor analyses of hands-on job sample tests and paper-and-pencil knowledge tests.

1. Basic Soldiering Skills (field techniques, weapons, navigate, customs and laws).
2. Safety/Survival (first aid, nuclear-biological-chemical safety).
3. Communications (radio operation).
4. Vehicle Maintenance.
5. Identify Friendly/Enemy Aircraft and Vehicles.
6. Technical Skills (specific to the job).

Table 3

Performance factors representing the common latent structure across all jobs in Project A sample. The criterion measures that comprise each factor are as indicated.

1) **Task Proficiency: MOS (Job) specific core technical skills**: The proficiency with which the individual performs the tasks which 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) **Task Proficiency: General or common skills**: In addition to the core technical content specific to an MOS, individuals in every MOS responsible for being able to perform a variety of general or common tasks --e.g., use of basic weapons, first aid, etc.. This factor represents proficiency on these general tasks.

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

3) **Peer 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. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient.

- Five scales from the Army-wide BARS rating form (gen. tech. performance, peer leadership, demonstrated effort, self development, gen. 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 (number of disciplinary actions and promotion 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.

TABLE 4

Measurement Methods Factors in Project A Job Performance Model

1) **Written Test Method** : That portion of the common variance among measures from the paper-and-pencil knowledge tests not predictable from (i.e., correlated with) any of the other related factor or performance construct scores.

2) **Ratings Method** : That portion of the common variance among measures from the rating instruments not predictable from (i.e., correlated with) any of the other related factor or performance construct scores.

PATTERNS IN SKILL LEVEL ONE PERFORMANCE IN REPRESENTATIVE ARMY JOBS: COMMON AND TECHNICAL TASK COMPARISONS

Roy C. Campbell
Charlotte H. Campbell
Earl L. Doyle

Human Resources Research Organization

Presented on Symposium,

"Job Performance: What Do Soldiers Know, What Can They Do?"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Patterns of Skill Level One Performance in Representative Army Jobs: Common and Technical Task Comparisons

Roy C. Campbell
Charlotte H. Campbell
and
Earl L. Doyle
Human Resources Research Organization

In the project for Improving the Selection, Classification and Utilization of Army Enlisted Personnel, commonly known as Project A, nine jobs or military occupational specialties (MOS) were covered intensively in the concurrent validation. The coverage included, among other measures, hands-on tests and written tests based on task samples for each MOS. The MOS, along with the number tested for each method, are shown in Table 1.

Table 1

MOS and Number Tested

MOS	SL1 Title	Written N	Hands-On N
11B	Infantryman	678	682
13B	Cannon Crewman	639	619
19E	Armor Crewman	459	474
31C	Single Channel Radio Operator	326	341
63B	Light Wheel Vehicle Mechanic	596	569
64C	Motor Transport Operator	668	640
71L	Administrative Specialist	501	494
91A	Medical Specialist	483	496
95B	Military Police	665	665

Army doctrine specifies that all skill level one soldiers are responsible for being able to perform all tasks in their MOS skill level one Soldier's Manual (SM) as well as the tasks listed in the skill level one Soldier's Manual of Common Tasks (SMCT). This latter document lists those tasks, known as Common Tasks, that every soldier, regardless of job or location, must be able to perform to survive in a hostile combat environment.

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

For Project A, the domain definition for each MOS consisted of these two types of tasks--those that were included because they were dictated by the soldier's job (MOS-specific or Technical tasks) and those that were included because Army doctrine requires all soldiers to perform minimum essential tasks dictated by exposure to wartime conditions (Common tasks). During the final process in which tasks from each domain were selected for testing, the process was structured so that the selection would represent the full range of task requirements in an MOS. Thus, for each MOS, the tasks tested include both Technical and Common tasks in both the hands-on and written components.

To be sure, the distinction between Technical and Common tasks is sometimes artificial. The skill level one soldier being trained probably does not discriminate between the two categories. And in many MOS, such as 11B, 95B, and 91A, there is little actual job distinction between MOS-specific and Common tasks. In these, and in some other MOS, if a task did not already exist in the SMCT, the job requirements would dictate the task be included as an MOS-specific task.

Yet much is made over Common Task requirements. The specific task concept for Common tasks began emerging in 1976 but is based on the long established Army tradition and concept that all soldiers, in combat, may be called upon to fulfill certain survival functions. The complexity of the modern battlefield has compounded, not diminished, this requirement. SMCT tasks receive as much attention and revision emphasis by TRADOC as do any of the MOS-specific technical tasks. Units are required to test selected common tasks annually. Army Training and Evaluation (ARTEP) and field exercises for all type units emphasize combat survival along with unit mission performance. But there are differences in emphasis as well. The 11B Infantryman literally lives with his M16 rifle; the 71L Administrative Specialist may only draw his/her M16 for maintenance and quarterly or semi-annual training. Yet by doctrine, each is equally responsible for certain M16 tasks. The question then, is whether there are distinctions among Army jobs in the performance of Common tasks and also whether there are significant distinctions between performance on Technical tasks and Common tasks. Project A, with the test results from over 5000 soldiers, provided an opportunity to examine this issue.

Method

In the 9 MOS, a total of 290 individual tasks were tested. Table 2 shows a breakout of these tasks by Technical and Common category and by test component. Almost all tasks tested in the hands-on component were also tested in the written component; however, there were some tasks tested by written component only.

Table 2

Distribution of Observations by Test Component

	Hands-On	Written
Technical	89	158
Common	60	123

The first analysis considered all the MOS combined (Table 3). There was only a slight and insignificant difference on hands-on results between the Common and Technical domains--the apparent difference being accounted for by the larger variance in performance in the Technical tasks. In the written tests, however, the difference in performance is significant, with higher performance levels reflected in the Common task performance. It should be noted however, that this difference may be the result of test difficulty. As yet, no overall item analysis of the written tests has been performed to identify difficulty patterns.

Table 3

Comparison of Technical and Common Task Performance on Hands-On and Written Tests For Nine MOS Combined

		Test Component	
Tasks		Hands-On	Written
Technical	N of Tasks	89	158
	Mean %	68.2	57.6
	S.D.	19.2	12.8
Common	N of Tasks	60	123
	Mean %	73.3	63.7
	S.D.	15.1	12.9
Test of Difference Between Common and Technical		$t = 1.721$ $p < .09$	$t = 3.948$ $p < .001$

Although the nine MOS were carefully selected to represent the entire domain of Army jobs, the Technical/Common tasks analysis continued by looking at the nine MOS broken down into families. These family classifications followed the groupings developed by McLaughlin, Rossmeissl, Wise, Brandt, and Wang (1984). Three families are represented: Family I is Combat (11B, 13B, 19E), Family II is Operations (31C, 63B, 64C), and Family III is Skilled/Technical (71L, 91A, 95B). (The 71L MOS actually belongs in a fourth job family--Clerical--but we have grouped it with the Skilled/Technical MOS for the analyses reported here.)

Table 4 shows the results by this family breakout. For the written tests there are no significant differences in performance between families, that is, where family membership affects outcome. In the hands-on tests, however, there appears to be a significant difference by family--Families I, II and III being each separated by about 5 points in performance. Closer examination however reveals that much of this difference by family is due to interaction between Common and Technical tasks within the family. Common task performance across families is quite consistent. The difference between families is accounted for almost solely by the Technical tasks, with 17 points difference in mean performance between the two most separated families.

Table 4

Performance Results Based on Family Membership

		Job Family		
		I - Combat	II - Operations	III - Skilled/ Technical
Hands-On Component Tasks				
Technical	N of Tasks	36	24	29
	Mean %	61.4	78.4	68.3
	S.D.	23.3	12.0	14.7
Common	N of Tasks	19	22	19
	Mean %	72.8	73.7	73.3
	S.D.	16.9	15.8	12.9
Written Component Tasks				
Technical	N of Tasks	63	49	46
	Mean %	55.1	58.1	60.4
	S.D.	13.0	11.2	13.0
Common	N of Tasks	44	39	40
	Mean %	63.0	63.0	64.2
	S.D.	12.9	13.7	12.3

Analysis of Variance: Job Family x Technical/Common

Hands-On Component

Source	SS	df	MS	F	p
Job Family	1910.16	2	955.08	3.28	.04
Technical/Common	543.78	1	543.78	1.86	.17
Family x Technical/Common	1561.08	2	780.54	2.68	.07
Error	41694.06	143	291.57		

Written Component

Source	SS	df	MS	F	p
Job Family	508.83	2	254.41	1.55	.21
Technical/Common	2316.63	1	2316.63	14.14	.00
Family x Technical/Common	205.18	2	102.59	.63	NS
Error	45062.95	275	163.86		

Within families however, there is always a significant difference between Technical task performance and Common task performance. However, this performance difference is not entirely consistent--in Families I and III, Common task performance is better than Technical performance. In Family II, the opposite is true for the hands-on test although the trend shown in Families I and III holds true for the written tests.

Conclusions

For a variety of reasons, relative differences in performance between Army jobs were expected. These differences can be variously attributed to innate task difficulty, assignments, training emphasis and even entrance requirements into the MOS. However it would appear that the Army policy regarding Common task proficiency appears to be working. While differences in performances between groups of MOS showed up as expected, these differences were almost entirely attributable to technical tasks within each group. Common task performance is remarkably uniform between Family groups. Based on Project A results it would appear the Army Common Task Management has produced its desired results.

References

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current and alternative ASVAB area composites, based on training and SQT information on FY1981 and FY1982 enlisted accessions (ARI Technical Report 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

EFFECTS OF TEST PROGRAMS ON TASK PROFICIENCY

Patrick Ford
R. Gene Hoffman

Human Resources Research Organization

Presented on Symposium,

"Job Performance: What Do Soldiers Know, What Can They Do?"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Effects of Test Programs on Task Proficiency

Patrick Ford and R. Gene Hoffman
Human Resources Research Organization

The general purpose of Project A is to predict job performance by establishing the relationship between entry measures and performance on a sample of job tasks in nine selected MOS (Eaton, Goer, Harris, & Zook, 1984). At a conceptual level the relation between applicants' ability and the tasks on a job ought to be stable so long as the job does not change. In practice, however, there are several mediators between ability and performance. Among the potential mediators are test programs that focus individual training in units. In these programs a central agency establishes a set of tasks that are to be tested and, presumably, trained in units. Data collected for Project A during June to November 1985 provide an opportunity to look at the effect of these programs on soldier performance.

This paper considers three programs that may affect task proficiency:

- Common Task Test (CTT). This is a hands-on test that all soldiers are to take each year. The Training and Doctrine Command selects a subset of tasks from the skill level one Soldier's Manual of Common Tasks. During the Project A data collection the operative CTT had 19 tasks. Across the nine MOS, the test samples for Project A included 14 of them. For comparison, 25 other non-CTT common tasks were also in the Project A data base.

- Expert Infantry Badge (EIB). This is a hands-on test that is administered to eligible infantrymen (MOS 11B). During the data collection it included 21 tasks of which 8 were included in the Project A 11B sample. The 11B test battery included 21 other tasks.

- Expert Field Medical Badge (EFMB). This is a written and hands-on test that is administered to medical specialists (MOS 91A). During the data collection the hands-on section included 32 tasks of which 10 were included in the Project A 91A sample. The 91A test battery included 20 other tasks.

The criterion measures for looking at the effect of the test programs are results from tests administered as part of Project A. There are two types of criterion measures:

- Hands-On Tests - These tests were based on direct observation of a soldier's performance of a job task. The tests were developed to provide consistent conditions for performance.

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Scores were percent of steps performed correctly or, in some cases, percent of product prepared correctly. There was a separate score for each task.

- Written Tests - These tests were in a multiple-choice format. Items were organized into subtests with each subtest corresponding to a job task. The score was percent correct by task.

During the data collection (which was the Project A Concurrent Validation), the tests had been administered to over 5000 skill level one (SL1) soldiers in nine MOS. The MOS covered along with the number of soldiers tested for each method are shown in Table 1.

Table 1.

MOS and Number Tested

MOS	SL1 Title	Written N	Hands-On N
11B	Infantryman	678	682
13B	Cannon Crewman	639	619
19E	Armor Crewman	459	474
31C	Single Channel Radio Operator	326	341
63B	Light Wheel Vehicle Mechanic	596	569
64C	Motor Transport Operator	668	640
71L	Administrative Specialist	501	494
91A	Medical Specialist	483	496
95B	Military Police	665	665

Approach

The CTT analyses were limited to the SL1 common tasks (defined as tasks included in the SL1 Soldier's Manual of Common Tasks). Performance on Project A tasks that were also on the CTT was compared with performance on Project A SL1 common tasks that were not on the CTT. The comparison was made on two levels--across all nine MOS and by MOS family. The MOS families were based on previous work (Rossmeissl, Wise, Brandt, & Wang, 1984) that identified four families: combat (11B, 13B, 19E); operations (31C, 63B, 64C); clerical (71L); and skilled technical (91A, 95B). The CTT analyses combined the clerical and skilled technical families. The analyses were conducted separately for hands-on and written criteria.

The analysis of specific MOS programs included all Project A tasks for MOS 11B and 91A respectively. Two comparisons per method were conducted for each program: (1) MOS program (EIB or EFMB) tasks and CTT tasks with Project A only tasks and (2) MOS program tasks with tasks not covered by the MOS program (including CTT tasks that were not in EIB or EFMB, respectively).

Results

The CTT comparisons are summarized in Table 2. Whether the differences are statistically significant depends on the orientation of the interpreter. If the question is simply "Does performance on this particular set of CTT tests differ from performance on this particular set of non-CTT tests?" essentially all of the differences would be statistically significant. That is, with test scores as percents, the extremely large number of soldiers tested yield standard errors of the mean for most tests at approximately .9. A more conservative standard is required, however, if the tests are treated as samples of their domain and the pertinent question is "Does performance on all tasks in the CTT domain differ from performance on all tasks in the non-CTT domain?" For the second question, the N is number of tasks sampled within task categories (e.g., CTT/Non-CTT) rather than soldiers.

The CTT comparisons were analyzed by means of a two way analysis of variance using tasks as subjects, with program membership as independent variables. Following the conservative interpretation (N as number of tasks), none of the differences are significant.

Table 2

Summary of Results on CTT Tasks and Non-CTT Common Tasks

Test Mode	Task Type	Family	N of Tasks	Mean	S.D.
Hands-On (60 cases)	CTT	All	28	76.73	8.31
		Combat	8	79.67	7.99
		Operations	11	74.92	7.48
		Skilled Tech. & Clerical	9	76.01	9.68
	Project A Common	All	32	70.40	18.84
		Combat	11	67.87	20.16
		Operations	11	72.49	21.61
		Skilled Tech. & Clerical	10	70.88	15.43
Written (123 cases)	CTT	All	56	65.77	12.54
		Combat	18	66.13	13.04
		Operations	18	67.93	12.34
		Skilled Tech. & Clerical	20	63.51	12.54
	Project A Common	All	67	61.91	13.01
		Combat	26	60.87	12.64
		Operations	21	60.37	14.20
		Skilled Tech. & Clerical	20	64.87	12.32

The EIB comparisons are summarized in Table 3. Here both hands-on comparisons are significant: Special program (EIB or CTT) with no special program ($F=6.022$, $P<.02$); and EIB with non-EIB ($F=6.21$, $P<.05$). Neither written comparison approaches significance.

Table 3

Summary of Results on 11B Special Program Tasks

Test Mode	Task Type	N of Tasks	Mean	S.D.
Hands-On (13 cases)	EIB & CTT	8	80.19	12.70
	Project A Only	5	57.26	16.50
	EIB	6	79.70	14.22
	Non-EIB	7	64.23	18.49
Written (28 cases)	EIB & CTT	13	61.58	9.51
	Project A Only	15	59.18	12.37
	EIB	8	62.21	9.80
	Non-EIB	20	60.03	11.69

The EFMB comparisons are summarized in Table 4. None of the differences are significant.

Table 4

Summary of Results on 91A Special Program Tasks

Test Mode	Task Type	N of Tasks	Mean	S.D.
Hands-On (16 cases)	EFMB & CTT	6	75.27	6.25
	Project A Only	10	70.58	12.49
	EFMB	5	76.43	6.24
	Non-EFMB	11	70.49	11.86
Written (30 cases)	EFMB & CTT	14	68.72	9.27
	Project A Only	16	65.66	13.06
	EFMB	11	70.33	9.40
	Non-EFMB	19	65.21	12.21

Discussion

Our reluctance to call the CTT differences significant ought not to be interpreted to mean that the CTT program makes no difference. All it means is that there is so much variation among the hands-on means for Project A only and so few cases overall that we can not say with confidence that hands-on performance on any set of tasks selected for CTT will be better than performance on tasks not selected. It is possible, for example, that some of the tasks not selected are more complex or require greater coordination than

tasks selected for CTT. Besides the possible sampling error among tasks we must also remember that the Project A results are a snapshot of a wide range of units at different points in their training cycles. Since the CTT effect could weaken over time, any evaluation that does not minimize the delay understates the effect.

The results do suggest that the portion of the CTT captured by Project A during the summer of 1985 had a positive association with hands-on scores. It is somewhat surprising that the difference was strongest in the MOS in the Combat family.

The EIB appears to be a very powerful program and must be considered when interpreting criterion data on 11B. It is less clear, however, that similar programs would achieve comparable results in any MOS. The impact of the EFMB on 91A, for example is not nearly as dramatic. Among the myriad of explanations for the difference, two seem to be especially appropriate. First, the EFMB has not had time to develop the credibility that the EIB has. The credibility of the program affects the number of people who are tested and, probably more important, the intensity of training that precedes the testing. Second, there may be a ceiling effect for 91A. Performance of medical specialists may be high enough without the program that any increment is small.

Conclusion

The impact of test programs on soldier performance is ambiguous. No program considered in this paper had a meaningful effect on performance as measured by written tests. The 1985 CTT apparently affected hands-on results in the Project A data but we cannot generalize that a comparable effect will occur every year. The effect was to equalize performance on a subset of common tasks across MOS mainly by increasing hands-on performance of soldiers in combat MOS. The EIB program had a strong effect on hands-on performance of infantrymen and should be considered as a moderator of 11B performance. However the EFMB program for medical specialists, though parallel to the EIB, did not have a comparable effect.

References

Eaton, N. K., Goer, M. H., Harris, J. H. and Zook, L. M. (Eds.). (1984). Improving the selection, classification, and utilization of army enlisted personnel: Annual report, FY1984 (ARI Technical Report 660). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current and alternative ASVAB area composites, based on training and SQT information on FY1981 and FY1982 enlisted accessions (ARI Technical Report 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

EFFECTS OF SOLDIER PERFORMANCE AND CHARACTERISTICS ON RELATIONSHIPS WITH SUPERIORS

Ilene F. Gast
Leonard A. White

U.S. Army Research Institute

Presented on Session, "Leadership"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Effects of Soldier Performance and Characteristics on Relationships with Superiors[1]

Ilene F. Gast and Leonard A. White
U.S. Army Research Institute for the Behavioral and Social Sciences

With the increasing emphasis on interactive leadership approaches (Jacobs, 1971; Graen 1976) has come a recognition of the contributions subordinates make to the leadership process. Although leaders may tend to have a characteristic style, they vary their behavior substantially in response to subordinate actions and needs. Graen has shown that leaders form different kinds of working relationships with their subordinates. Relationships range from "in-group" ones characterized by mutual support and trust to "*out-group*" *ones* where both parties do only what is required by the formal employment contract.

Graen (1976) notes that relationships formed early in one's career have lasting effects. Based on a longitudinal investigation of management trainees, Wakaybayashi and Graen (1984) conclude that a newcomer's relationship with his or her superior serves motivating and mentoring functions that help the newcomer to assimilate into the organization and to gain access to information and resources central to the functioning of the work unit. This experience gives newcomers the confidence they need to set higher performance goals. Thus, the relationships that first tour soldiers form with their superiors are not only important from the standpoint of socialization into the Army, but may also affect career progression, and leadership potential.

Past research has shown that subordinates' performance is a powerful determinant of subsequent treatment by superiors (e.g., Greene, 1975). Generally, poor performers are more likely to have low quality relationships with their superiors. However, because this phenomena has been investigated primarily in the laboratory, field research is needed.

There also is evidence that relatively stable personal dispositions enable some subordinates to form more positive relationships with superiors. Graen and his associates (Graen, Novak, & Sommerkamp, 1982) demonstrated the importance of subordinates' growth need strength to the formation of effective relationships with superiors. However, with the exception of Graen's work and research by Hough, Gast, White and McCloy (1986), researchers have not adequately adressed the potential effects of individual differences on subordinates' interactions with their superiors. Such research is needed.

Using data from Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel (Eaton, Goer, Harris & Zook, 1984), this paper examines how working relationships between superiors and subordinates are directly affected by subordinates' job performance, temperament and ability. In addition, this paper explores non-linear effects of soldier ability and temperament on working relationships with superiors.

Method

Subjects

Subjects were 5,123 first term soldiers in 9 military occupational specialties (MOS): 683 infantrymen (11B), 636 cannon crew members (13B), 489

---

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

tank crew members (19E), 349 radio teletype operators (31C), 618 light wheel vehicle mechanics (63B), 670 motor transport operators (64C), 502 administrative specialists (71L), 487 medical specialists (91A) and 689 military police (95B). Within the sample, 88% of the soldiers were male and 12% female. Of those who reported their racial origin, 23% were black, 3% were hispanic, 70% were white, and 4% replied "other". On the average, soldiers had been in the Army for 18 months and with their present companies for about a year. To facilitate data analysis, jobs were grouped into four occupational clusters identified by McLaughlin, Rossmeissl, Wise, Brandt and Wang (1982). The Combat cluster included MOS 11B, 13B and 19E; MOS 31C, 63B and 64C comprised the Operations cluster; MOS 71L made up the Clerical cluster and the remaining MOS, 91A and 95B comprised the Skilled Technical cluster.

Instruments

Supervisor Behavior Questionnaire. The authors wrote items to tap categories of supervisory activities identified through analysis of 400 behavioral examples of effective and ineffective leadership. These items required subjects rate statements about their supervisor using a 5-point scale from Very Seldom or Never (1) to Very Often or Always (5). The resulting questionnaire was field tested in a sample of 696 first term enlisted (White, Gast, & Rumsey, 1985) and revised prior to administration in the present sample. Principal factor analysis with promax rotation revealed five factors with eigenvalues greater than one: Inspiration/Support, Participation, Structuring Work, Fairness/Discipline, Work Allocation. The present research employed only the scales corresponding to the first two factors; these scales were most similar to scales measuring qualities of "in group" relationships in previous research (Vecchio & Gobdel, 1984; Novak, 1985). Typical items on the 9-item Inspiration/Support scale included "Your supervisor understands your problems and needs" and "Your supervisor wants to make you give your best effort". The 4-item Participation scale contained items like "You are permitted to use your own judgment in solving problems". Reliabilities (Chronbach's alpha) for these two scales were .82 and .70 respectively.

General cognitive ability. The Armed Services Vocational Aptitude Battery (ASVAB) was administered to all subjects prior to entering military service. A composite of four ASVAB subtests, known as the Armed Forces Qualification Test (AFQT), served as the measure of general cognitive ability.

Temperament. Hough, Kamp and Barge (1984) developed ten scales to assess temperament constructs shown to be related to criteria of work performance in previous studies. The resulting inventory, Assessment of Background and Life Experiences (ABLE), was tested on 470 soldiers at three forts. These data guided revisions to the items and scales. When subjected to principal factor analysis with varimax rotation, the revised scales yielded three factors with eigenvalues greater than one: Dependability, Achievement Orientation, and Emotional Stability. Scales measuring self-esteem, dominance, energy level, and work orientation comprise the Achievement Orientation factor. The Emotional Stability factor assesses the degree of stability vs. reactivity of emotions. The Dependability factor includes measures of conscientiousness, non-delinquency, support for rules and regulations, and respect for traditional values. Factor scores for these three scales used in the analyses.

Job knowledge tests. Through job analysis important knowledge areas were identified for each MOS. Project A personnel, assisted by subject matter specialists, developed items to tap these knowledges. The overall job knowledge test score was the percentage of items answered correctly by each soldier.

Hands-on task proficiency tests. Critical tasks were identified to represent the task domain for each MOS. A multiple step proficiency test was developed for each task, and each step was scored pass or fail. For each task, the score was the proportion of steps passed; then these task scores were averaged to yield an overall hands-on test score (Campbell, Campbell, Rumsey & Edwards, 1985).

Army-wide performance rating scales. Eleven 7-point behaviorally anchored rating scales were developed to assess soldier effectiveness across army jobs. These scales went beyond task performance to include aspects of socialization and commitment to the organization. Ten scales covered specific aspects of soldier effectiveness; the eleventh scale required an assessment of overall effectiveness. Supervisors' ratings on this eleventh scale were employed in the present analyses (Campbell et al., 1985).

Procedure

After receiving training in the use of the behavior anchored rating scales, supervisors, in groups of 3-15, evaluated their subordinates. The mean number of supervisors providing the ratings for each ratee ranged from 1.66 to 1.83. Ratings were averaged across raters to form an overall Army-wide effectiveness rating for each ratee. Tests of job knowledge and hands-on task proficiency were also administered to the soldiers.

Results and Discussion

The performance measures (i.e., hands-on, job knowledge and supervisory ratings) were standardized within each MOS cluster. Then, moderated regression techniques were used to examine determinants of leadership within each MOS cluster. The "moderating" effect of one independent variable on another is indicated by a significant increase in explained variance due to entry of the cross-product term after all main effects have been entered entered into the model. Separate models were constructed for Supportive and for Participatory leadership. Explanatory variables were entered into the equations in sets; models were tested in the following sequence: (a) main effects of individual difference variables, (b) main effects of performance variables, (c) all main effects, (d) all main effects and interactions between ability and temperament variables, (e) all main effects and interactions among temperament variables, and (f) all main effects and all interactions.

Table 1 summarizes the results from all models tested. Among the performance measures, supervisors' assessments of subordinate performance predicts reported leadership most consistently. Looking across all of the models, work sample performance and job knowledges do not contribute significantly to the prediction of Supportive leadership. Task proficiency predicts participation within the Operations Operations and Skilled Technical MOS clusters; job knowledge predicts participation within the Operations cluster.

Independently, the set temperament of variables accounts for at least as much variance in reported leadership as individual differences in job performance do. When considered apart from the performance measures, the three temperament measures are significant predictors of rated leadership across MOS. Combined with the performance measures, the independent contribution of the temperament measures weakens somewhat, suggesting that these measures share variance with supervisory ratings. Cognitive ability is a significant predictor in only one MOS cluster, the combat related jobs. Although ability does not generally make a direct contribution to the prediction of rated

leadership, past research suggests that it is an antecedent of job knowledge which, in turn affects hands-on performance and supervisors' assessments of subordinate performance (White, Borman, Hough & Hoffman, 1986). Given the mediational role that job knowledge plays, its failure to to have a direct effect on reported leadership in the present investigation is not surprising.

Without exception, the combination of individual differences in temperament and job performance performance variables accounts for more variance in leadership measures than either set of variables considered alone. However, the addition of interaction terms offers little advantage. In no case do they increase the amount of variance accounted for by more than two percentage points. Further, the interactions have no consistent pattern of significance.

Finally, the two leadership variabiles appear to differ in hov well they can be predicted by the independent variables. With the exception of the Combat job cluster, regardless of which model is tested, the independent variables account for more variance in Participation than Inspiration/Support. Further, Achievement Orientation is a significant predictor of Participation, but not of Supportive leadership. Thus, in determining the amount of support a leader will provide, subordinate attributes may contribute less than other determinants of leadership behavior (e.g., leader attributes, organizational norms and values, resource allocation), whereas in most MOS, participation may depend more heavily on subordinate characteristics.

In summary, soldiers who report receiving higher levels of support from their superiors tend receive higher scores in dependability and emotional stability and are seen by their superiors as effective performers. Further, soldiers who report more involvement in work related decisions have the preceding characteristics and are also scored as more achievement oriented.

The present research successfully extended past research in two important ways. First, it demonstrated in a field setting that performance predicts reported leadership. Second, although performance affects soldiers' treatment by their superiors, individual differences in job-related temperament factors are at least equally important. Further, both sets of variables make independent contributions to the prediction of reported leadership. Because treatment by superiors can be predicted from relatively stable individual differences, supervisory treatement should be expected to generalize across supervisors and through time. Thus, subordinates who negotiate more effective relationships with their superiors during the first tour should be expected do so throughout their careers. Additionally, it is likely that future bosses will see these individuals as more effective.

Future research should trace the careers of individuals in the Project A database to determine if, in fact these predictions hold. Further, the present research assumed one-way causality; future research might address the bi-directional causality of superior-subordinate interactions.

References

Campbell, C. H. Campbell, R. C., Rumsey, M. G., & Edwards, D.C. (October, 1984). Development and Field Test of Task-Based MOS-Specific Criterion Measures. (Technical Report No. 717). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (October, 1984). Improving The Selection Classification and Utilization of Army Enlisted Personnel: Annual Report, 1984 Fiscal Year. (Technical Report No. 660). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Graen, G. B. (1976). Role-making processes within complex organizations. In M. Dunnette (Ed.), Handbook of industrial organizational psychology. Chicago: Rand McNally.

Graen, G. B., Novak, M. A., & Summerkamp, P. (1982). The effects of leader-member exchange and job design on productivity and satisfaction: Testing a dual attachment model. Organizational Behavior and Human Performance, 30, 109-131.

Greene, C. N. (1975). The reciprocal nature of influence between leader and subordinate. Journal of Applied Psychology, 60, 187-193.

Hough, L. M., Kamp, J. D., & Barge, B. A., (1984). Utility of Temperament, Biodata, and Interest Assessment for Predicting Job Performance: A Review and Integration of the Literature. Minneapolis: Personnel Decisions Research Institute.

Hough, L. M., Gast, I. F., White, L. A., & McCloy, R. (1986, August). The Relation of Leadership and Individual Differences to Job Performance. Paper Presented at the Meeting of the American Psychological Association, Washington, D. C.

Jacobs, T. O. (1971). Leadership and Exchange in Formal Organizations. Alexandria, VA: Human Resources Research Organization (HumRRO).

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current armed services vocational aptitude battery (ASVAB) composites. (Technical Report No. 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Novak, M. A. (1985). A study of leader resources as determinants of leader-member exchange. Doctoral dissertation, University of Cincinnati, 1984). Ann Arbor, MI: University Microfilms International.

Vecchio, R. P., & Gobdel, B. C., (1985). The vertical dyad linkage model of leadership: Problems and prospects. Organizational Behavior and Human Performance, 34, 5-20.

Wakabayashi, M., & Graen, G. B., (1984). The Japanese career progress study: A 7-year follow-up. Journal of Applied Psychology, 69, 603-614.

White, L. A., Gast, I. F., & Rumsey, M. G., (1985). Leader behavior and the performance of first term soldiers. Paper presented at the meeting of the Military Testing Association, San Diego, CA.

White, L. A., Borman, W. C., Hough, L. M., & Hoffman, R. G. (1986, August). A path analytic model of job performance ratings. Paper presented at the meeting of the American Psychological Association, Washington, D. C.

Table 1

Results of Regression Analyses for Each MOS Cluster

MODEL	Main Effects AFQT 1	ACH 2	DEP 3	EMOT 4	HO 5	JK 6	SR 7
					Inspiration/Support		
Clerical MOS							
INSP = IND. DIF.	NS	*	**	*			
INSP = PERF.					NS	NS	**
INSP = PERF + IND. DIF.	NS	NS	**	NS	NS	NS	**
INSP = MAIN EFF. + AFQT*TEMP	NS	NS	NS	NS	NS	NS	**
INSP = MAIN EFF. + TEMP*TEMP	NS	NS	**	NS	NS	NS	**
INSP = MAIN EFF. + ALL INTERACTIONS	NS	*	NS	NS	NS	NS	**
Combat MOS							
INSP = IND. DIF.	NS	NS	**	**			
INSP = PERF.					NS	NS	**
INSP = PERF + IND. DIF.	NS	NS	**	**	NS	NS	**
INSP = MAIN EFF. + AFQT*TEMP	**	NS	**	**	NS	NS	**
INSP = MAIN EFF. + TEMP*TEMP	NS	NS	**	**	NS	NS	**
INSP = MAIN EFF. + ALL INTERACTIONS	*	NS	**	**	NS	NS	**
Operations MOS							
INSP = IND. DIF.	NS	**	**	**			
INSP = PERF.					NS	NS	**
INSP = PERF + IND. DIF.	NS	**	**	**	NS	NS	**
INSP = MAIN EFF. + AFQT*TEMP	NS	NS	**	**	**	NS	**
INSP = MAIN EFF. + TEMP*TEMP	NS	**	**	**	NS	NS	**
INSP = MAIN EFF. + ALL INTERACTIONS	NS	NS	**	**	NS	NS	**
Skilled Technical MOS							
INSP = IND. DIF.	NS	*	**	**			
INSP = PERF.					NS	NS	**
INSP = PERF + IND. DIF.	NS	NS	**	**	NS	NS	**
INSP = MAIN EFF. + AFQT*TEMP	NS	NS	NS	NS	NS	NS	**
INSP = MAIN EFF. + TEMP*TEMP	NS	NS	**	**	NS	NS	**
INSP = MAIN EFF. + ALL INTERACTIONS	NS	NS	NS	NS	NS	NS	**
	1	2	3	4	5	6	7

					Participation		
rical MOS							
RT = IND. DIF.	NS	**	**	**			
RT = PERF.					NS	**	**
RT = PERF + IND. DIF.	NS	**	*	**	NS	NS	*
RT = MAIN EFF. + AFQT*TEMP	NS	*	NS	NS	NS	NS	NS
RT = MAIN EFF. + TEMP*TEMP	NS	**	**	**	NS	NS	*
RT = MAIN EFF. + ALL INTERACTIONS	NS	*	NS	NS	NS	NS	NS
abat MOS							
RT = IND. DIF.	NS	**	**	**			
RT = PERF.					NS	NS	**
RT = PERF + IND. DIF.	NS	**	**	*	NS	NS	**
RT = MAIN EFF. + AFQT*TEMP	NS	**	**	NS	NS	NS	**
RT = MAIN EFF. + TEMP*TEMP	NS	**	**	*	NS	NS	**
RT = MAIN EFF. + ALL INTERACTIONS	NS	**	**	NS	NS	NS	**
erations MOS							
RT = IND. DIF.	NS	**	**	**			
RT = PERF.					**	NS	**
RT = PERF + IND. DIF.	NS	**	**	**	**	**	**
RT = MAIN EFF. + AFQT*TEMP	NS	**	NS	**	**	**	**
RT = MAIN EFF. + TEMP*TEMP	NS	**	*	**	**	**	**
RT = MAIN EFF. + ALL INTERACTIONS	NS	**	NS	**	**	**	**
illed Technical MOS							
RT = IND. DIF.	NS	**	**	**			
RT = PERF.					*	NS	**
RT = PERF + IND. DIF.	NS	**	**	**	*	NS	**
RT = MAIN EFF. + AFQT*TEMP	NS	NS	NS	NS	**	NS	**
RT = MAIN EFF. + TEMP*TEMP	NS	**	**	**	**	NS	**
RT = MAIN EFF. + ALL INTERACTIONS	NS	NS	NS	**	NS	NS	**

THE PROJECT A CONCURRENT VALIDATION DATA COLLECTION

James H. Harris
John P. Campbell
Charlotte Campbell

Human Resources Research Organization

Presented on Symposium,

"Project A Concurrent Validation: Preliminary Results"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

The Project A Concurrent Validation Data Collection 1,2

James H. Harris
Human Resources Research Organization

John P. Campbell
University of Minnesota

Charlotte H. Campbell
Human Resources Research Organization

Introduction

The purpose of this paper is to describe the Project A concurrent validation data collection and relate some "lessons learned" about the administration of large scale data collections. During this data collection, predictor and criterion measures were administered to approximately 9,500 entry-level soldiers and rating scales were administered to approximately 7,000 supervisors of these soldiers. The original Project A Research Plan specified a concurrent validation target sample size of 600-700 skill level (SL1) job incumbents for each of 19 mos, using procedures that had been tried out and refined during the predictor and criterion field tests. The Research Plan specified 13 data collection sites in the United States (CONUS) and two in Europe (USAEUR). the number of sites was the maximum that could be visited within the Project's budget constraints, which dictated that sites be chosen to maximize the probability of obtaining the required sample sizes. the data collection schedule, by site, is shown in Figure 1.

The basic sampling plan, data collection team training, data collection procedures, and lessons learned are presented in the following sections.

Sampling Plan

The general sampling plan was to use the Army's World-Wide Locator System to identify all the first-term enlisted personnel in the 19 mos at each chosen site who entered the Army between 1 July 1983 and 30 July 1984. If possible

[1]This research was funded by the U.S. Army research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

[2]The material in this paper is from two sources: Campbell, C.H., & Hoffman R.G. (in press). Concurrent validation hands-on data collection: Lessons learned. Alexandria, VA: Human Resources Research Organization (HumRRO).
Human Resources research Organization, American Institutes for Research, Personnel Decisions Research Institute and Army Research Institute (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report. ARI Technical Report 746. Alexandria, VA: Army Research Institute.

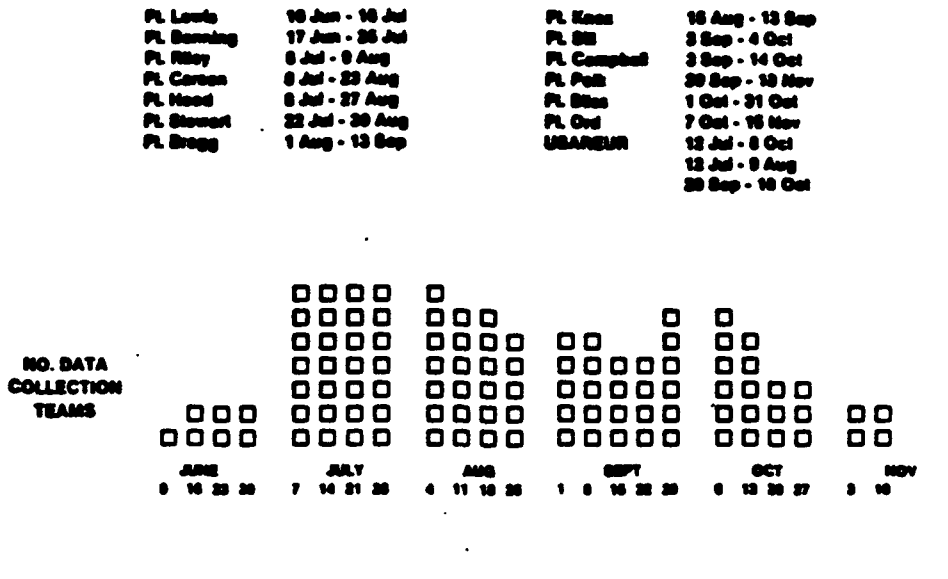


Figure 1. Concurrent validation schedule.

the individual's unit identification was also to be retained. The steps described below were then followed. The intent was to be as representative as possible while preserving enough cases within units to provide a "within rater" variance estimate for the supervisor and peer ratings.

A. Preliminary Steps

1. Identify the subset of MOS (within the sample of 19) for which it would be possible to actually sample people within units at specific posts. That is, given the entry date "window" and given that only 50-75 percent of the people on <u>any</u> list of potential subjects could actually be found and tested, what MOS are large enough to permit sampling to actually occur? List them.

2. For each MOS in the subset of MOS for which sampling is possible, identify the smallest "unit" from which 6-10 people can be drawn. Ideally, we would like to sample 4-6 units from each post and 6-12 people from each unit. For the total concurrent sample this would provide enough units to average out or account for differential training effects and leadership climates, while still providing sufficient degrees of freedom for investigating within-group effects such as rater differences in performance appraisal.

3. For the four MOS in the Preliminary Battery (PB) sample, identify the members of the PB sample who are on each post.

B. The ideal implementation would be to obtain the Alpha Roster list of the total population of people at each post who are in the 19 MOS and who fit our "window." The lists would be sent to the data collection manager where the following steps would be carried out.

1. For each MOS, randomize units and randomize names within units.

2. Select a sample of units at random. The number would be large enough to allow for some units being truly unobtainable at the time of testing.

3. Instruct the Point-of-Contact (POC) at the post to obtain the required number of people by starting at the top of the list and working down (as in the Batch A field test) within each of the designated units. If an entire unit is unavailable, go on to the next one on the list.

4. In those MOS for which unit sampling is not possible, create a randomized list of everyone on the post who fits the window. Instruct the POC to obtain the required number by going down the list from top to bottom (as in the Batch A field tests).

C. If it is not possible to bring the Alpha Roster to the data collection manager, provide project staff at the post to assist the POC in carrying out the above steps.

1. If it is not possible to randomize names at the post, first use the World-Wide Locator to obtain a randomized list, carry the list to the post and use it to sample names from units drawn from a randomized list of units. If there are only 6-8 units on the post, then no sampling of units is possible. Use them all.

D. If it is not possible for project personnel to visit the post, then provide the randomized World-Wide Locator list to the POC and ask him or her to follow the sampling plan described above with written and telephone assistance. That is, the POC would identify a sample of units (for those MOS for which this is possible), match the unit roster with the randomized World-Wide Locator list, and proceed down each unit until the required number of people was obtained. If the POC can generate their own randomized list from the Alpha Roster, so much the better. The World-Wide Locator serves only to specify an a priori randomized list for the POC.

E. If none of the above options is possible, then present the POC with the sampling plan and instruct him or her to obtain the required number of people in the most representative way possible (the Batch B procedure).

The final sample sizes are shown by post and by MOS in Figure 2. Note that it was not always possible in all MOS to find as many as 600 incumbents with the appropriate accession dates at the 15 sites. Some MOS simply aren't that big.

Data Collection Team Training

Each data collection team was composed of a Test Site Manager (TSM) and six or seven project staff members who were responsible for test and rating scale administration. The teams were made up of a combination of regular

	Batch A MOS									Batch Z MOS										
Location	11B	13B	19E	31C	63B	64C	71L	91A	95B	12B	16S	27E	51B	54E	55B	67N	76W	76Y	94B	Total	% Total
Fort Benning	65	23	61	7	33	39	16	9	33	13	15	3	0	12	18	9	13	15	12	316	1.95
Fort Bliss	0	20	30	15	61	65	17	0	44	15	5	2	0	14	0	12	6	31	30	367	3.68
Fort Bragg	68	46	0	0	37	25	41	10	72	82	75	13	19	72	20	7	62	39	62	730	7.74
Fort Campbell	90	28	0	20	60	65	54	64	43	90	23	10	0	32	16	42	91	61	66	757	8.03
Fort Carson	60	50	77	30	49	53	30	33	46	49	57	13	0	25	7	0	23	60	67	689	7.31
Fort Hood	26	56	0	30	60	28	38	50	60	51	60	4	12	62	36	66	72	61	57	767	8.13
Fort Knox	29	32	111	16	38	48	22	65	31	63	10	6	0	8	12	0	20	29	34	524	5.56
Fort Lewis	75	46	13	11	43	46	23	27	56	27	25	1	11	51	31	20	68	41	56	631	6.69
Fort Ord	30	0	0	14	30	62	32	43	51	51	7	0	1	6	7	15	23	60	28	425	4.51
Fort Polk	73	47	19	29	47	47	18	46	44	60	45	9	0	16	7	23	26	51	35	443	6.87
Fort Riley	30	63	55	27	26	45	35	30	40	31	20	0	0	25	52	0	20	39	45	579	6.14
Fort Sill	0	198	0	20	43	51	44	0	29	43	11	0	0	0	0	15	7	35	32	437	4.63
Fort Stewart	44	46	39	17	28	51	31	45	65	30	39	9	0	17	29	26	44	34	35	617	6.54
USAREUR	132	122	130	130	122	121	116	119	118	120	78	61	41	96	54	63	105	134	133	1963	20.80
Total	702	667	503	366	637	686	514	501	692	706	470	167	108	436	291	276	490	630	612	9430	
% Total	7.44	7.07	5.33	3.88	6.76	7.27	5.45	5.31	7.34	7.47	4.98	1.56	1.15	4.60	3.09	2.93	5.20	6.68	6.49		

Figure 2. Concurrent validation sample soldiers by MOS by location.

project staff and individuals (e.g., graduate students) specifically recruited for the data collection effort. The test site manager was an "old hand" who had participated heavily in the field tests. This team was assisted by eight NCO scorers (for the hands-on tests), one company-grade officer POC, and up to five NCO support personnel, all recruited from the post.

The project data collection teams were given three days of training at a central location. During this period, Project A was explained in detail, including its operational and scientific objectives. After the logistics of how the team would operate (transportation, meals, etc.) were discussed, the procedures for data entry from the field to the computer file were explained in some detail. Every effort was made to reduce data entry errors at the outset via correct recording of responses and correct identification of answer sheets and disketttes.

Next, each predictor and criterion measure was examined and explained. The trainees took each predictor test, worked through samples of the knowledge tests, and role played the part of a rater. Considerable time was spent on the nature of the rating scales, rating errors, rater training, and the procedures to be used for administering the ratings. All administrative manuals, which had been prepared in advance, were studied and pilot tested, role playing exercises were conducted, and hands-on instruction for maintenance of the computerized test equipment was given.

The intent was that by the end of the three-day session each team member would (a) be thoroughly familiar with all predictor tests and performance measures, (b) understand the goals of the data collection and the procedure

for avoiding negative critical incidents, (c) have had an opportunity to practice administering the instruments and to receive feedback, and (d) be committed to making the data collection as error-free as possible.

As noted above, eight NCO scorers were required for Hands-On test scoring. They were recruited and trained using procedures very similar to those used at each post in the criterion field tests. Training took place over one full day and consisted of (a) a thorough briefing on Project A, (b) an opportunity to take the tests themselves, (c) a check-out of the specified equipment, and (d) multiple practice trials in scoring each task, with feedback from the project staff. The intent was to develop high agreement for the precise responses that would be scored as GO or NO-GO on each step.

Data Collection Procedure

The data collection proceeded as follows: The first day was devoted to equipment and classroom set-up, general orientation to the data collection environment, and a training and orientation session for the post POC and the NCO support personnel.

On the first day of actual data collection the soldiers who arrived at the test site were divided randomly into two equal groups, identified as Group 1 or 2. Each group was directed to the appropriate area to begin the administration for that group. They rotated under the direction of the test site manager through the appropriate block according to the schedule.

For soldiers in a Batch Z MOS, like 12B, the procedure took one day. For soldiers in a Batch A MOS, like MOS 91A, the procedure was similar but took two days to rotate the soldiers through the appropriate blocks. The measures administered in each block are shown in Figure 3.

BATCH A MOS 4 Blocks 4 Hrs. Each		BATCH Z MOS 2 Blocks 4 Hrs. Each	
Block 1	Predictor Tests	Block 1	Predictor Tests
Block 2	School and Job Knowledge Tests Army-Wide Ratings	Block 2	School and Job Knowledge Tests Army-Wide Ratings
Block 3	MOS Specific Hands-On Tests		
Block 4	MOS Ratings MOS Specific Written Tests		

Figure 3. Concurrent validation test outline.

Lessons Learned

Collecting data from 16,000 soldiers in 15 locations over six months is a difficult task, one that requires careful planning, attention to detail, an ability to adapt, a fondness for crisis management, and a special relationship with the telephone. For anyone planning an effort of like grandeur (or even grander), a few lessons learned from some of the survivors seems appropriate. We divide the lessons into three categories: planning, coordinating, and operating Each category is briefly discussed below.

Planning. Start as early as possible (18 months before collecting data) to identify the support you will need, to include personnel, equipment, facilities, and time requirements. Once you know what you need and when you need it, schedule a series of briefings with the Commanders. Start at the top with the CG of FORSCOM, TRADOC, and USAREUR and work your way through a series of briefings until you reach the local POC responsible for seeing that you get what you need when you need it. Be prepared to change your plans at each step to meet local concerns. Once you meet and brief your POC, you can begin coordinating.

Coordinating. The closer the time to begin data collecting, the more frequently you will speak to the POC. Expect to speak daily when you get within 30 days of data collection. In some instances, you may have to make a trip to the installation for a final coordination meeting. Be prepared to be very flexible with regard to the installation's internal schedule.

Operating. Most of the lessons learned in this category have to do with hands-on testing.

1. Many instances of equipment variation can be (and were) anticipated. Test developers and site coordinators must find out what major pieces of equipment are not likely to be available at the selected sites in advance of actual testing if high quality tracked tests are to be prepared.

2. Printed scoresheets must be proofed carefully to ensure that for every step which should be scored, a score can be recorded.

3. Scorers must be thoroughly trained, not only on how to set up and administer the tests, but also on how to record data on the scoresheets. They must be given practice in using the scoresheets (not just talked through it) before testing, and monitored closely during testing, especially with the first few soldiers tested. Continual monitoring must also occur throughout the testing.

4. Scorers and hands-on managers must document meticulously who was tested on what, and also who wasn't tested on what, and why.

5. Experienced hands-on managers are often able to implement procedures to deal with equipment malfunctions or variations, but these too must be documented.

6. Completed scoresheets must be checked as soon as possible after testing so that careless or incorrect scoring can be detected, and the errant scorer can be retrained.

POST DIFFERENCES IN HANDS-ON TASK TESTS

R. Gene Hoffman

Human Resources Research Organization

Presented on Session, "Issues in Hands-On Performance Testing"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Post Differences in Hands-On Task Tests

R. Gene Hoffman
Human Resources Research Organization

One of the major efforts for the U.S. Army's Selection and Classification Project (Project A) has been the development of hands-on performance measures. The effort required preparation of tests to cover approximately 15 tasks for soldiers in nine different job specialties (MOS). Because of equipment differences within certain MOS, it was necessary to create alternate versions of some tests. Thus, 103 different task tests were prepared. Eleven tests were used in more than one MOS with the number of tests per MOS ranging from 14 to 27. As part of the concurrent validation data collection effort, these tests were administered during 1985 to approximately 500 to 600 soldiers per MOS. In order to collect that volume of data, test sites included 13 different Army posts in the United States plus European test sites. At the European sites, approximately 120 soldiers for each MOS were tested. At the CONUS sites, the numbers of soldiers per MOS per site ranged from 9 to 110 with typical numbers being near 30, near 45 or near 60 because of scheduling requirements. The tests were administered in blocks of two to four tasks per test station with typically one NCO in the respective MOS at each site handling test administration for all soldiers at any given station.

Given these "road show" requirements for data collection, considerable effort was made to standardize the hands-on testing procedures. These efforts included attention to test set-up and scoring instructions and to the training of test administrators. Prior to concurrent data collection, test procedures were pilot tested on a small sample of soldiers using four to five test administrators and then field tested on approximately 150 soldiers. Administrator training included five phases: (1) presentation of general testing principles, (2) familiarization with individual test station requirements, (3) practice, (4) review by contractor personnel prior to data collection, and (5) monitoring by contractor personnel during data collection. Further details concerning test construction and administration are presented in Campbell et al., (1985) and Campbell (in preparation).

Given that hands-on testing has a history of being susceptible to scorer differences (e.g., Maier, 1983), this paper examines differences between posts in hands-on test scores and the extent to which any such post differences are not "real" differences, but are, in some way, artifacts of the measurement process. Thus, analyses examined alternative sources of variance in hands-on test scores that could account for any mean differences between posts. Candidate measures for explaining differences available in the Project A data set include: (1) written tests, (2) supervisor and peer ratings of performance, (3) practice, (4) time in service, and (5) ability. Post effects were estimated after variance due to these measures was removed from the hands-on tests (using hierarchical multiple regression) and compared with post effects prior to any adjustment.

Analysis

Analyses were conducted for every hands-on test in all nine MOS. No adjustment was made for tests appearing in more than one MOS. That is, repeated tests were treated as separate observations. Thus, there were 147 observations of post differences where an observation is a test/MOS combination. The first series of analyses estimated unadjusted post effects (percent of variance in hands-on score accounted for by post alone) and post effects adjusted for written test scores (except, obviously, those tasks tested only in the hands-on mode), task ratings by peers and by supervisors, overall performance ratings by peers and supervisors, practice (composite of self ratings of recency and frequency of task performance), time in service (test date minus entry date), and general ability (AFQT). In conducting these analyses, significant reductions in sample sizes between post only and adjusted post analyses were observed for all MOS. The reductions were most attributable to missing ratings. Therefore, an alternative or "reduced" adjustment model was also examined in which ratings were excluded. Thus, for each of the 147 tasks, three different R^2s were calculated between post and hands-on scores: (1) an unadjusted "post alone" R^2, (2) an adjusted R^2 for post after all other variables in the "full model" were controlled, and (3) an adjusted R^2 for post after all other variables in the "reduced (ratings excluded) model" were controlled. Adjusted R^2s were calculated as the increase in R^2 when post was added after all control variables in a hierarchical multiple regression predicting hands-on score. Mean sample sizes for these analyses were 500.21 for post alone, 164.01 for the "full model" (i.e., all variables) and 341.77 for the "reduced model."

The R^2s between post alone and hands-on scores estimate the extent of between post differences in hands-on scores. These were compared to the R^2s for post and hands-on scores after variance due to the other variables in the full and reduced models were controlled. Differences in variance accounted for by post (i.e., differences in R^2s) were calculated as indices of the bias resulting from post differences. Thus, two bias indices for each hands-on test resulted from these analyses: a "full model" bias and a "reduced model" bias. The term bias has been used in response to the question: "Would standardizing hands-on scores by post bias those scores?" Positive values for these bias indices would suggest that any post differences are to some extent real and that standardizing would introduce bias. On the other hand, near zero values suggest that post differences are unrelated to other measurements of performance, therefore may reflect measurement error, and that standardization may be justified.

The above analyses were conducted on a task by task basis. From these analyses it is not possible to tell whether the "post" effects are actually at the post level or are more correctly attributable to scorer differences. Two approaches were used to address this question, neither of which is definitive. First, if "post" effects (within an MOS) were operating consistently for all tasks within an MOS (e.g., motivational differences between posts), then it should be possible to account for post variance in any one task by removing variance associated with the hands-on test scores for other tasks within each MOS. Thus for each task, an adjusted R^2 for post effects were examined after variance associated with other MOS tasks was removed. An "other tasks" bias index was constructed as the difference

between post effects alone and this "other tasks" adjusted R^2. If this index is near zero, the "post" effects are task specific and not consistent across tasks within an MOS.

A second way to partially dissect the task by task post effects is to examine scorer-within-post variance capitalizing on the instances where two or more scorers scored the same test at the same post either by general design (i.e., duplicate equipment and test stations in the test plan) or by local variation (i.e., an early finishing scorer helping at another station).

The series of analyses examining post effects controlling for performance on other hands-on tests occurred some time after the first, and in that interval two 91A tracked tests were merged; therefore 146 separate tasks were analyzed. Again a "bias" variable was calculated as the difference between post effects alone and adjusted post effects.

Results

Results for these analyses are summarized in Table 1 below. All data points were either R^2s (for the Post Only analyses), increases in R^2s (for the full, reduced and other task model analyses), or differences between R^2s (for the bias variables). Thus, table entries are the means, standard deviations, minimums and maximums for these R^2s across the 147 tasks.

Uncorrected post differences account for an average of 19% of the variance in hands-on test scores, indicating the presence of post differences in hands-on scores. Post effects range from 2% to 50%. For only 36 of the 147 tasks is the post effect less than 10% of the hands-on variance. Furthermore, there is no evidence that post differences can be consistently attributed to written test scores, practice, ratings, ability, or time in service. Mean bias from the full and reduced model analyses are both very near zero suggesting that removing post differences by standardization would not bias the hands-on scores.

Table 1

Hands-On Test Variance (R^2) Associated With Post With and Without Controls and Associated Adjustment Bias

	Variance Associated with Post				Standardization Bias		
	Post Only Model	Full Model	Reduced Model	Other Tasks Model	Full Model Bias	Reduced Model Bias	Other Task Model Bias
Mean R^2	0.19	0.22	0.18	0.12	-0.03	0.01	0.07
S.D. R^2	0.11	0.11	0.11	0.08	0.08	0.05	0.06
Min. R^2	0.02	0.01	0.01	0.01	-0.25	-0.24	-0.04
Max. R^2	0.50	0.52	0.52	0.34	0.24	0.23	0.33

Results for the "other tasks" model are presented in Table 1. Bias as estimated by this model is somewhat larger than the others and suggests that to some extent post differences for any given task are related to post differences for other tasks. However, certainly not all of the task level post effects are explained.

Table 2 indicates that the 147 tasks are rather homogeneous with regard to reduced model bias. For the 147 tasks, 114 reduced model bias indices are between -.05 and .05. The other bias indices are similarly homogeneous. Thus, the post effects that are present remain so after attempts to explain them are considered and that trend is consistent across all tasks.

Table 2

Distribution of Reduced Model Bias Across 147 Hands-On Tests

Reduced Model Bias	Frequency	Percent
-0.30>		
-0.25>	0	.00
-0.20>	1	.68
-0.15>	0	.00
-0.10>	4	2.72
-0.05>	6	4.08
-0.00>	56	38.10
0.05>	58	39.46
0.10>	18	12.24
0.15>	1	.68
0.20>	2	1.36
0.25	1	.68

The final analysis made use of the duplication of scorers for some tasks at some posts. Because this duplication was not systematically planned, some instances of duplication of scorers were due to a scorer at one post scoring only one or two soldiers. Such cases are not very illuminating. To avoid them, only tasks for which degrees of freedom for scorers-within-post was at least 5 were examined. Forty tasks met this criterion (degrees of freedom ranged from 5 to 23). For these tasks, the mean scorers-within-post effect accounted for 4.6% of the hands-on variance. This number probably underestimates the size of the scorer effect because post effects were still confounded by scorer effects. That is, for all but a few tasks in this analysis, several posts were represented by only one scorer. For the thirteen tasks with 10 or more degrees of freedom for scorers within post (and fewer posts with only one scorer), 6.4% of the hands-on variance is associated with scorer differences. While it is not possible to totally disentangle post versus scorer differences, it is probably safe to conclude that there were consistent scorer differences, and that some of the differences among posts are attributable to scorer differences.

These analyses unfortunately are like trying to show that something does not exist when we can look in only so many places. That is, we are trying to

rule out alternative explanations for the post effects while we are limited in the availability of ways to look. Given the evidence, unwanted post effects at the task level can not be ruled out, and the standardization of hands-on test means by post appears justified.

One may wonder what might be the negative consequences if the decision to standardize by post is incorrect. The most damaging consequence would be an introduction of error leading to a reduction in the predictability of hands-on measures. To shed some light on this possibility, the predictability of standardized and unstandardized hands-on test scores were compared using the reduced model variables (i.e. R^2s for predicting hands-on tests from written tests, experience, practice, time, and ability). Across the 147, the average difference between the two R^2 is .02 with the standardized hands-on scores being slightly less predictable. The standard deviation of the difference across the 147 tasks is .05. Thus, across the tasks standardizing has little effect one way or the other on the predictability of the hands-on scores.

Summary

In summary, post effects on hands-on scores were present and no alternative explanation of those effects was found. This leaves the implication that the post differences reflect error in the measurement process. Second, the post effects seem to be operating idiosyncratically at the task level, i.e., as the post or scorer effects unique to each task, rather than as the post level effects consistent for all tasks in an MOS. Third, while it is not possible to totally disentangle post and scorer, some of the between post differences are probably due to scorer differences. Fourth, post differences should be controlled in further statistical analyses of hands-on test scores. And finally, even if this conclusion is incorrect, statistical corrected by standardizing by post will not have a grave impact on the predictability of the hands-on scores.

References

Campbell, C. H. (1986). Developing basic criterion scores for hands-on tests, job knowledge tests, and task rating scales (In preparation). Alexandria, VA: Human Resources Research Organization.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., and Edwards, D. C. (1985). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Maier, M. H. (1983). Using job performance tests as criteria for validating qualifications standards (Memorandum CNA 83-3123.09). Alexandria, VA: Center for Naval Analyses.

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

ESTIMATES OF TASK PARAMETERS FOR TEST AND TRAINING DEVELOPMENT

R. Gene Hoffman
Patrick Ford

Human Resources Research Organization

Presented on Session, "Issues in Job Analysis"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Estimates of Task Parameters for Test and Training Development

R. Gene Hoffman and Patrick Ford
Human Resources Research Organization

The Army's Project A is a large scale effort to validate the ASVAB and a battery of new selection and classification tests for enlisted soldiers. The effort requires comprehensive job performance measures as validation criteria. In the early stages of the project the domains of nine selected MOS were described to allow the selection of performance variables which could be translated into reliable and representative samples of those performance domains. The problem was to narrow down large domains. The problem is a familiar one in the military context in both the testing and training arenas. That is, job analyses have already been conducted and doctrinal directive written which specify at great length the tasks which soldiers in each MOS are supposed to be able to perform. Far too many tasks are designated as part of the job than any particular training or testing program can cover.

To reduce the task domains for Project A, five task parameters were identified as potentially significant for the selection of sets of representative tasks. These include (1) the relative importance among the tasks, (2) the similarities among the tasks, (3) the performance frequency of each task, (4) the difficulty of each task, and (5) the variability in performance for each task. Details concerning all of these parameters and how they were used in task selection is reported elsewhere (HumRRO & AIR, 1984) and will not be repeated. Our focus is retrospective. Performance measures have been constructed and administered to approximately 400 to 650 soldiers in each of the nine MOS. This provides the opportunity to examine the validity of the task selection data for three of the task parameters: (1) task difficulty, (2) task variability, and (3) task frequency.

Data Base

The "population" for this analysis is tasks rather than people, and the sample is the overlap between the set of tasks for which hands-on performance tests were administered during Project A's concurrent validation phase and the AOSP task list as refined for task selection uses (Campbell, et al., 1985). Some adjustments were necessary because equipment variation necessitated the use of alternative test forms whereas AOSP statements were equipment generic. Thus, 135 tasks spanning the nine MOS were included in the analysis.

Difficulty and variability task parameters were estimated during task selection using a single rating scale. For each AOSP task within their respective MOS, subject matter experts (SME; Ns ranged from 10 to 26 for the nine MOS) were asked to describe the performance distribution of soldiers. They were asked to indicate: "Out of 10 soldiers, how many can do the task: (1) All of the time?, (2) Most of the time?, (3) About half of the time?, (4) less than half of the time?, or (5) Never?" SMEs were also given an escape option of "Not observed." Each set of SME responses therefore represented a

frequency distribution of task performance. By assigning performance values (1 to 5) to the response intervals, a performance mean and standard deviation was computed for each task for each SME. For each task, these individual SME means and standard deviations were averaged across SME, excluding SME who responded with "not observed." Thus, the average SME mean and average SME standard deviation became the difficulty and variability parameters used in the task selection process. Interrater reliabilities within each MOS were in the .70s and .80s for task difficulty and in the .50s and .60s for task variability for the nine MOS. SME (generally E-6 to E-7) rated approximately 150 to 300 tasks within their MOS. Further details are presented in HumRRO and AIR (1984).

Task frequency data used in task selection were taken directly from the AOSP survey results for skill level one soldiers. The specific index was the percent of soldiers reporting that they performed each task.

On the criterion side of this validation, actual test statistics from the concurrent validation data collection provide task difficulty and variability estimates. Performance on these tasks was assessed using four modes: (1) hands-on tests, (2) written tests, (3) peer ratings and (4) supervisor ratings. Means and standard deviations for all four measurement modes were used as criteria against which SME derived estimates were compared. Hands-on and written test scores were percent correct for either steps or items. Performance ratings were given by both peers and supervisors on a 7-point scale ranging from "among the very worst" to "among the very best" at the end points with "about the same as others" at the midpoint.

Project A concurrent validation also included a job history questionnaire completed by each soldier. For each task in the hands-on test sample, the questionnaire asked soldiers to describe on a five point scale how recently they had performed the task and how frequently in the past six months they had performed the task. These responses, averaged across soldiers, provide an independent assessment of task experience for validating AOSP frequency data.

Convergence between task selection data and concurrent validation measurement data was assessed with simple correlations. Correlations within each MOS and across all MOS are reported.

For MOS level correlations for task difficulty and variability estimates, Ns range from 13 to 17 tasks for hands-on and ratings measures, and 12 to 16 for written measures. Not all MOS had the same number of hands-on tests and for six tasks there was no matching written test. One task had no matching rating. Across the nine MOS, the total numbers of tasks were 135 for correlations involving hands-on data, 129 for correlations involving written tests and 134 for correlations involving ratings. Since AOSP frequency data were not available for all tasks, MOS level correlations of task experience were based on Ns which ranged from 10 to 15, with a total of 108 tasks across all MOS.

Results

Table 1 presents correlations between SME estimates and data-based estimates of task difficulty. At the MOS level the correlations fluctuate from -.04 to .95 and given the small Ns on which these correlations are computed such large fluctuations are expected. Confidence interval estimates depend on sample size, size of the observed correlation and are not symmetrical. For simplicity however, it is useful to use one central confidence interval for reviewing a set of correlations. Thus, the 95 percent confidence interval, using the lowest N (12) and an average $\underline{r}$ near .50 is $\underline{r}$ = -.10 to $\underline{r}$ = .84 which is not very different from the range observed in Table 1. Across all MOS, SME ratings of task difficulty are more predictive of rating means as given by peer and supervisors than written and hands-on test score means. The .95 confidence interval for total sample correlations using the lowest N (129 for written tests) and an average $\underline{r}$ = .50 is $\underline{r}$ = .36 to $\underline{r}$ = .62. Thus, the variation among the correlations is not greater than chance.

Table 1

Correlations Across Tasks Between SME Means and Measurement Mode Means For Each MOS and Total Sample

MOS	Hands On	Written	Peer Rating	Sup. Rating
11B	0.50	0.21	0.69	0.80
13B	0.92	0.70	0.81	0.82
19E	0.54	0.47	0.95	0.93
31C	0.58	0.13	0.83	0.86
63B	-0.04	0.07	0.69	0.56
64C	0.34	0.51	0.49	0.65
71L	0.71	0.66	0.36	0.30
91A	0.30	-0.11	0.65	0.74
95B	0.21	0.15	0.29	0.31
TOTAL	0.43	0.33	0.59	0.62

Table 2

Correlations Across Tasks Between SME Standard Deviations and Measurement Mode Standard Deviations For Each MOS and Total Sample

MOS	Hands On	Written	Peer Rating	Sup. Rating
11B	0.62	0.34	0.86	0.68
13B	0.75	0.37	0.77	0.77
19E	0.51	0.52	0.28	0.17
31C	0.60	0.28	0.17	0.54
63B	0.16	0.14	0.07	-0.02
64C	0.26	0.12	0.87	0.82
71L	0.22	0.39	0.70	0.30
91A	0.25	0.06	0.18	0.68
95B	0.50	0.39	0.33	0.48
TOTAL	0.35	0.26	0.42	0.48

Table 2 presents the analogous correlations between SME estimates of task variability and data based estimates. Again at the MOS level the correlations fluctuate from -.02 to .86. Again, however correlations do vary more than expected by chance.

For reference, intercorrelations among task means and among task standard deviations are presented in Tables 4 and 5 in an Appendix.

Table 3 presents correlations between Project A frequency and recency and AOSP task experience estimates, as well as correlations between an unweighted linear composite of the frequency and recency with AOSP frequency.

Looking at the composite, correlations range from .02 to .90 for the within MOS data (.95 confidence interval for an average $\bar{r}$ = .56 is $\bar{r}$ = -.10 to $\bar{r}$ = .88). Across all MOS, frequency and recency means for the 108 tasks each correlate .46 with AOSP frequency (.95 confidence interval is $\bar{r}$ = .31 to $\bar{r}$ = .58). Frequency and recency means correlated .91 with each other, so that using a composite of the two does little to strengthen the relationship between the two sets of experience data.

Table 3

Correlations Across Tasks Between AOSP Frequencies and Job History Responses for each MOS and Total Sample

MOS	Frequency	Recency	Composite
11B	0.85	0.90	0.88
13B	0.55	0.46	0.52
19E	0.53	0.43	0.50
31C	0.14	0.09	0.13
63B	0.00	0.05	0.02
64C	0.65	0.81	0.76
71L	0.11	-0.08	0.02
91A	0.88	0.92	0.90
95B	0.49	0.58	0.53
TOTAL	0.46	0.46	0.47

Discussion

Results indicate that, in the absence of hard performance data, SME estimates can provide reasonably valid, though certainly not perfect, estimates of difficulty and variance. Given validity coefficiencies in the .40 to .60 range, SME estimates of task difficulty can be useful for making gross judgments differentiating particularly hard or easy tasks. In essence, that was the use made of the SME difficulty estimates during task selection with the very hard and the very easy tasks generally not selected for testing. Thus, there is some degree of range restriction in the SME ratings used in the present analysis and the validity of the SME estimates may be understated.

The strength of the relationship between SME task difficulty and performance rating means is interesting in light of the performance rating scale. Theoretically the scale should have led to means near the mid-point for every task, with near zero variance across tasks. Realistically, our knowledge of common rating errors led us to hedge our bets here. Thus, we analyzed the performance rating means expecting to find convergence with SME means. Even though the standard deviation across tasks of the rating means were restricted to .28 and .36 for peers and supervisors, respectively, the variance in task means that did exist was strongly associated with SME task

difficulty estimates. Raters apparently had a hard time making purely normative judgments. That is, raters may have been reluctant to give average or below average ratings on tasks that almost all soldiers perform well.

Validities for the SME estimates of performance variability are lower. Intercorrelations among all estimates of task variability show a similar reduction (compare Tables 4 and 5 in the Appendix). Thus, relative differences among tasks in variance seem more affected by test mode than do their relative differences in difficulty. This makes SME estimates of task performance variability less useful for task selection.

Project A and AOSP estimates of task frequency show modest but perhaps more limited convergence than might be expected from two self-reports of essentially the same phenomenon: participation in various tasks. There are, however, several differences between the two which may have reduced their convergence. First, they provide different experience indices (percent of soldiers who do a task from AOSP data versus average number of times a task is done from Project A data) which may have distorted the relative distributions for tasks done as a daily part of the job (e.g., type a DF for 71L clerks) versus tasks practiced only during set training periods (e.g., load, reduce a stoppage and clear an M16). Second, the surveys were conducted at different times (several years apart for some MOS), and any instability over the intervening time periods would reduce convergence. This was the case for two MOS with low experience convergence (31C and 63B) where preparation of task tests was more cumbersome than other MOS because of the variety and continuing evolution of equipment. Finally, AOSP estimates were based on a sample of the entire first tour, while Project A estimates were based on soldiers representing a more limited range of one to two years time in service. As soldiers increase in time in service, their job duties may expand and change. The distinctions between the two surveys are important caveats for interpreting either set of experience data.

References

Campbell, C. H., Campbell, R. C., Rumsey, M. G., and Edwards, D. C. (1985). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization (HumRRO) and American Institutes for Research (AIR) (1984). Selecting job tasks for criterion tests of MOS proficiency (ARI Working Paper RS-WP-84-25). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Appendix

Table 4

Intercorrelations Among Measurement Mode Means

Mode	Hands-On	Written	Peer Rating	Supervisor Rating
Hands-On	1.00			
Written	0.52			
Peer Rating	0.58	0.40	1.00	
Supervisor Rating	0.53	0.37	0.93	1.00

Table 5

Intercorrelations Among Measurement Mode Standard Deviations

Mode	Hands-On	Written	Peer Rating	Supervisor Rating
Hands-On	1.00			
Written	0.40			
Peer Rating	0.48	0.17	1.00	
Supervisor Rating	0.42	0.20	0.70	1.00

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

USING CONFIRMATORY FACTOR ANALYSIS TO AID IN ASSESSING TASK PERFORMANCE

Jeffrey J. McHenry
American Institutes for Research

James H. Harris
Human Resources Research Organization

Scott M. Oppler
American Institutes for Research

Presented on Symposium,

"Innovations in Manpower Research Methods:
Current Practice and Suggestions"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

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

Jeffrey J. McHenry
American Institutes for Research
Washington, DC

James H. Harris
Human Resources Research Organization
Alexandria, VA

Scott M. Oppler
American Institutes for Research
Washington, DC

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

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

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

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

---

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

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

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

Method

Subjects

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

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

Measures

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

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

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

Results

Model of Task Performance

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

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

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

Table 1

Measurement of Task Factors by MOS

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

Analyses

The objective of this study was to test whether the observed

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

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

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

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

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

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

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

Table 2

Fit between the Task Factor Model and the Observed Correlations

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

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

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

Discussion

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

INFLUENCE OF ENVIRONMENT, ABILITY, AND TEMPERAMENT ON PERFORMANCE IN ARMY MOS

Darlene M. Olson
U.S. Army Research Institute

Walter C. Borman
Personnel Decisions Research Institute

Presented on Session, "Organizational Effectiveness"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

INFLUENCE OF ENVIRONMENT, ABILITY AND TEMPERAMENT ON PERFORMANCE IN ARMY MOS

Darlene M. Olson
U.S. Army Research Institute[1]

Walter C. Borman
Personnel Decisions Research Institute

Job performance has been conceptualized as a product of abilities, skills, and personal characteristics that individuals bring to the Army, of environmental experiences that influence a soldier after enlistment, and of the person's motivation to perform. Although a substantial portion of the total variability in performance criteria can be explained by individual difference factors, work environment variables related to support, training opportunities, and perceived job importance have been found to have weak, but consistently significant relationships with supervisory ratings of soldier effectiveness, Army-wide rating factors (e.g., Personal Discipline) and measures of hands-on task proficiency (Olson & Borman, 1986).

The impact of cognitive abilities, temperament, work environment and their possible interactive effects on job performance should be investigated. Peters & O'Connor (1980) have proposed that environmental factors may moderate the relationships between ability and performance. In contrast, Schmidt and Hunter (1977) have contended that the prediction of performance from ability is stable across situations and over time for various jobs. More current research (e.g., Staw & Ross, 1985) has found dispositional effects for job satisfaction criteria. Hence, research suggests that both person and environment factors should play a role in explaining the variability in soldier performance.

The model of soldier effectiveness advanced here assumes that performance is influenced by a soldier's abilities and temperament, which are measured when entering the military, and individual perceptions of the work environment developed through experience with the Army job setting. In this context, the purpose of this research was to investigate potential moderating effects of work environment dimensions on the relationship between individual differences and job performance in four clusters of Army jobs.

Method

Subjects. The sample contained 5080 first-term Army enlisted personnel in 9 different jobs. There were 673 infantrymen, 629 cannon crewmen, 485 armor crewmen, 351 radio operators, 618 light-wheel vehicle mechanics, 659 motor transport operators, 500 administrative specialists, 485 medical specialists, and 680 military police. These MOS were sampled at 11 continental United States and four European Army installations. These jobs were grouped into one combat (11B, 13B, and 19E MOS) and three non-combat clusters [Clerical (71L MOS), Operations (31C, 63B, and 64C MOS), and Skilled Technical (91A and 95B MOS)]. Previous empirical research (McLaughlin, et. al., 1984) demonstrated that the above clusters are sufficient to group Army jobs on the basis of aptitudes measured by ASVAB.

Performance Measures. Criterion development work was conducted by the Project A contractors and included construction of the following measures: 1) Army-wide rating scales relevant for evaluating soldiers in any first-tour

---

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

Army job, 2) job-specific rating scales, 3) hands-on task proficiency measures, and 4) job knowledge tests. The Army-wide rating scales were developed using a variant of the behaviorally-anchored rating scale methodology, and emphasize performance dimensions relevant to any MOS (e.g., maintaining equipment). The job-specific scales, which were also 7-point behavior summary scales, focus on narrow performance areas relevant to a designated job (e.g., transporting personnel for the motor transport operator job). The hands-on tests consisted of 15 MOS-specific tasks. Hands-on scores were computed for each soldier by averaging the proportions passed across the tasks tested. Multiple choice tests were developed to assess job knowledge relevant to important and representative tasks in an MOS. A total job knowledge score for each research participant was derived as a percentage of the number of items answered correctly. Factor-analysis of the performance ratings resulted in an interpretable solution: 1) Effort and Leadership 2) Personal Discipline and 3) Military Bearing (Campbell, Hanser, & Wise, 1986). Factor scores for the performance ratings, along with an overall soldier effectiveness composite based on the unit weighting of ratings on the Army-wide dimensions were used in subsequent analyses.

Work Environment Measures. The Army Work Environment Questionnaire (AWEQ), a revised 53 item multiple choice questionnaire measures the following Army environmental constructs: 1) Resources, 2) Supervisor Support, 3) Training/Opportunities to Use MOS skills, 4) Job/Task Importance, and 5) Cohesion/Peer Support. AWEQ items are answered using a 5-point frequency rating scale (e.g., 1 = Very Seldom or Never to 5 = Very Often or Always). Respondents are asked to indicate how often each environmental situation described in an item occurs on their present job. Items consist of statements such as "You get recognition from supervisors for the work you do" (Supervisor Support). Five standardized unit weighted factor scores are derived for the AWEQ.

Cognitive Ability. A composite measure of four subtests from the Armed Services Vocational Aptitude Battery (ASVAB), known as the Armed Services Qualifications Test (AFQT), was used as an assessment of general cognitive abilities.

Temperament Measures. The Assessment of Background and Life Experiences (ABLE) inventory (Peterson, Hough, Ashworth, & Toquam, 1986), which includes ten tempera-ment/biodata scales, was administered as a self-report measure of soldier temperament. From factor analysis of the ABLE a three factor solution emerged: 1) Achievement, 2) Dependability and 3) Adjustment. The Achievement factor has items loading from the Self-Esteem, Work Orientation, Dominance and Energy-Level scales. The Dependability factor contains items from the Non-delinquency, Traditional Values, Conscientiousness, Cooperativeness, and Internal Control scales. The Adjustment factor has items loading from the Emotional Stability scale.

Procedures. The rating scales were administered to groups of 15 or fewer peers or supervisors of the target ratees after they were trained using a combination error and accuracy training program. During the peer rating sessions, raters (who were in addition ratees and members of the research sample) also responded to the AWEQ. The ABLE inventory was administered in separate small group sessions. Task proficiency measures were admininstered to each soldier by experienced job incumbents or supervisors, who were trained to evaluate and score each hands-on task. MOS-specific job knowledge tests were given to groups of 15-30 soldiers.

Results

Regression Analyses. Moderated regression analysis was used to estimate the relationships of ability, temperament, perceptions of the work environment, and their interactions to typical performance ratings and more objective performance criteria. A series of four separate regression models were built for each of the four performance measures nested in each job cluster. First, the separate performance variables were regressed on an individual differences model, which contained AFQT and three temperament factor scores to determine the contribution of individual differences at the time of enlistment to subsequent job performance. Second, an environmental model, which contained the five work environment constructs was used to predict the separate performance measures to examine the amount of variance explained by these variables. Third, a full model containing both individual differences and environmental factors was tested. Finally, a set of interactions among the predictors (ability X temperament, ability X environment, and temperament X environment factors) was added to the full model and the separate performance criteria were regressed on it to determine the post-enlistment interrelationships among environmental/organizational influences on soldier performance and the expression of individual differences in ability and temperament on the job.

The regression analyses are presented in Table 1. In each of the four job clusters, the highest multiple correlations were observed for the prediction of job knowledge, with R ranging from .37 to .57, $p < .05$. Ability explained the largest amount of variance in job knowledge scores. Generally, the full model of individual differences accounted for more variance in the performance measures than was explained by the environmental model. However, for both the Operations and Skilled Technical job clusters, higher multiple correlations (Rs = .32 and .26, respectively) were obtained for the prediction of task proficiency from environmental models as compared with the individual differences model (R = .14 and .20, respectively).

In the clerical and combat jobs, soldier ability and temperament characterized by Dependability accounted for the most variance in the performance criteria. For the Operations and Skilled Technical MOS, the temperament factors (particularly Dependability and Achievement) explained significant variability in the rating measures, and soldier ability tended to account for significant variance in the more objective performance measures. The environmental model accounted for 3-10% of the variability in criterion measures for the separate job clusters. The largest standardized regression coefficients were observed for the prediction of ratings from Supervisor Support and Job/Task Importance factors. Training had a strong main effect for the prediction of task proficiency and job knowledge measures for the MOS clusters. Further, those variables with the largest standardized beta coefficients in the separate individual differences and environmental models were retained in the full model of main effects for the clusters.

Table 2 shows that ability X Job/Task Importance interaction effects tended to be significant across MOS clusters (except for Operations) and performance measures (except for hands-on). The majority of interaction effects were concentrated between individual differences related to soldier temperament and work environment constructs. Specifically, temperament factors related to Dependability and Adjustment interacted with soldier perceptions of Job/Task Importance, level of Supervisor Support, and available

Table 1

Standardized Regression Coefficients in the Multiple Regression Models for the ROS Clusters

Clusters	1	2	3	4	5	6	7	8	9	R^2	$R^2_{Adj.}$	F
1. INDIVIDUAL DIFFERENCES[1]												
Effectiveness	.12*	.19**	.14**	.11*						.09	.08	.28
Discipline	.12*	.06	.17**	.12*						.07	.05	.26
Hands-On	.34**	0	.16**	.08						.16	.15	.48
Job Knowledge	.31**	.02	.16**	.06						.22	.21	.57
2. ENVIRONMENT[2]												
Effectiveness					-.06	.27**	-.05	.08	-.05	.07	.06	.26
Discipline					-.03	.23**	-.04	.07	0	.07	.06	.26
Hands-On					-.01	.06	.21**	.06	.01	.06	.05	.24
Job Knowledge					-.02	.07	.11*	.08	.05	.04	.03	.20
3. INDIVIDUAL DIFFERENCES + ENVIRONMENT												
Effectiveness	.12*	.15**	.10*	.09	-.04	.25**	-.04	.01	-.06	.13	.11	.36
Discipline	.11*	.02	.14**	.10	-.04	.23**	-.05	.01	-.01	.12	.09	.35
Hands-On	.33**	.01	.14**	.07	-.02	.06	.19**	-.02	-.02	.20	.18	.45
Job Knowledge	.31**	.02	.16**	.06	-.03	.06	.10*	-.02	.01	.23	.22	.37

Combat	1	2	3	4	5	6	7	8	9	R^2	$R^2_{Adj.}$	F
1. INDIVIDUAL DIFFERENCES[1]												
Effectiveness	.11**	.15**	.21**	.14**						.11	.10	.33
Discipline	.13**	-.01	.23**	.16**						.15	.14	.39
Hands-On	.21**	.08**	-.07*	-.01						.06	.05	.24
Job Knowledge	.46**	.02	.10**	.05						.23	.24	.50
2. ENVIRONMENT[2]												
Effectiveness					-.08*	.17**	0	.09**	.05	.05	.05	.22
Discipline					-.04	.28**	-.05	.11**	0	.08	.06	.24
Hands-On					-.14**	.01	.30**	.03	.01	.05	.04	.22
Job Knowledge					-.12**	.03	.16**	.09**	.03	.03	.02	.17
3. INDIVIDUAL DIFFERENCES + ENVIRONMENT												
Effectiveness	.11**	.13**	.17**	.12**	-.07*	.12**	.02	.04	.02	.12	.12	.35
Discipline	.14**	-.03	.29**	.13**	-.06	.14**	-.03	.07*	-.02	.17	.16	.41
Hands-On	.23**	.06*	-.06**	-.02	-.10**	0	.22**	.05	0	.11	.10	.33
Job Knowledge	.36**	0	.08**	.04	-.09**	-.02	.14**	.12**	0	.28	.27	.53

Operations	1	2	3	4	5	6	7	8	9	R^2	$R^2_{Adj.}$	F
1. INDIVIDUAL DIFFERENCES[1]												
Effectiveness	.04	.13**	.22**	.08**						.07	.07	.26
Discipline	.03	.01	.25**	.11**						.09	.09	.30
Hands-On	.15**	0	.04	-.02						.02	.02	.14
Job Knowledge	.39**	-.05	.09**	.02						.16	.16	.40
2. ENVIRONMENT[2]												
Effectiveness					-.06*	.10*	0	.16**	0	.04	.04	.20
Discipline					-.01	.13**	-.02	.10**	0	.04	.04	.20
Hands-On					-.12**	.05	.30**	.21**	.08**	.10	.09	.32
Job Knowledge					-.13**	-.05	.16**	.14**	.14**	.07	.07	.26
3. INDIVIDUAL DIFFERENCES + ENVIRONMENT												
Effectiveness	.03	.10**	.20**	.06*	-.06*	.04	.01	.12**	0	.09	.08	.30
Discipline	.03	-.02	.25**	.09**	-.03	.10**	-.01	.08*	0	.11	.10	.33
Hands-On	.14**	-.04	.02	-.05	-.11**	-.05	.20**	.22**	.08**	.12	.11	.35
Job Knowledge	.30**	-.06**	.09**	.01	-.13**	-.07	.15**	.15**	.12**	.22	.22	.47

Skilled Technical	1	2	3	4	5	6	7	8	9	R^2	$R^2_{Adj.}$	F
1. INDIVIDUAL DIFFERENCES[1]												
Effectiveness	.03	.22**	.22**	.11**						.11	.11	.33
Discipline	.03	-.02	.28**	.13**						.10	.10	.32
Hands-On	.15**	.03	.05	.04						.04	.04	.20
Job Knowledge	.30**	-.01	.15**	.06						.11	.11	.33
2. ENVIRONMENT[2]												
Effectiveness					-.10**	.16**	.09*	.16**	.05	.10	.09	.32
Discipline					-.09*	.21**	.12**	.05	.05	.09	.08	.30
Hands-On					-.12**	-.02	.26**	0	.03	.07	.06	.26
Job Knowledge					-.01	.11*	-.15**	-.04	.05	.03	.02	.17
3. INDIVIDUAL DIFFERENCES + ENVIRONMENT												
Effectiveness	.03	.15**	.17**	.09**	-.09*	.14**	.11**	.09*	.04	.17	.16	.41
Discipline	.03	-.04	.24**	.12**	-.12**	.17**	.13**	.03	.04	.16	.15	.40
Hands-On	.19**	.02	.05	.04	-.12**	-.03	.27**	-.01	.02	.11	.10	.33
Job Knowledge	.29**	0	.14**	.06	-.02	.08	-.14**	-.06	.04	.14	.13	.37

Note. [1]Individual differences model contains the set of temperament factors and ability.
[2]The Environment model consists of the five unit environment factors.
Variables in the regression models are 1=Ability, 2=Achievement, 3=Dependability, 4=Adjustment, 5=Concern, 6=Supervisor Support, 7=Training, 8=Job/Task Importance, and 9=Cohesion/Peer Support.
*$p < .05$, **$p < .01$.

Table 2

Summary of Significant Interactions Among Ability, Temperament, and Work Environment in the Prediction of Performance

Interactions	Overall Effectiveness			Personal Discipline		Hands-On				Job Knowledge		
	CL	CO	ST	CL	CO	CL	CO	OP	ST	CL	CO	OP
Ability X Achievement							A					
Ability X Dependability	A				A							
Ability X Adjustment								A				
Ability X Resources								B				
Ability X Support			B									
Ability X Job Importance	A		A	A							B	
Achieve X Support	A											
Achieve X Training							A					
Depend X Resources						B						B
Depend X Support		A										
Depend X Job Importance						B					A	
Depend X Cohesion/Peer Support												B
Adjust X Resources								A				A
Adjust X Support	A							A				
Adjust X Job Importance		B			A							
Adjust X Cohesion/Peer Support									A			

Note. The job clusters are CL = Clerical, CO = Combat, OP = Operations, and ST = Skilled Technical. Significant interaction effects were not found for the rating criteria in the Operations job cluster. Significant interaction effects were not found for the Personal Discipline rating factor and job knowledge test for the Skilled Technical job cluster. Significant interactions are A = $p < .05$; B = $p < .01$.

organizational Resources. Fewer significant interactions were observed between cognitive ability (AFQT) and temperament in the prediction of job performance. Training X Achievement and Cohesion/Peer Support X Adjustment interactions significantly predicted task proficiency in Combat and Operations clusters respectively. Further, for the Operations jobs, several significant interaction effects between soldier perceptions of Resources and individual differences were found to predict maximal performance criteria.

Generally, when designated interactions are added to the full model of main effects, only about 1% of the variance in performance beyond that explained by main effects can be attributed to interactions. However, for the Clerical MOS, interaction effects accounted for an additional 3-7% of the variability in soldier performance, with higher percentages of explained variance associated with the more objective performance criteria.

Discussion

This research examined relationships among individual differences in ability and temperament, perceptions of the Army work environment, and the performance of first term enlisted personnel. Findings revealed that individual differences and environmental perceptions have independent effects on performance in the four job clusters. Some differential effects were found across job clusters with maximal performance (e.g., job knowledge and task proficiency) predicted best from cognitive ability (AFQT) in the Clerical and Combat jobs.

Significant effects for the work environment indicate that both types of typical performance ratings are predicted from the more climate-oriented constructs of Supervisor Support and Job/Task Importance; particularly in the

Combat and Operations clusters. In contrast, soldiers' perceptions of Training and their opportunities to utilize MOS skills, as well as the availability of Resources (e.g., tools and equipment) tended to predict both job knowledge and task proficiency measures for all job groups. Interaction results show that both temperament and work environment factors moderate the relationship between ability and performance. In addition, work environment factors related primarily to Supervisor Support, Resources, and Job/Task Importance, and to a lesser extent Training tended to moderate the relationships between individual temperament factors and performance.

These findings tentatively indicate that job performance is influenced not only by individual differences in ability, but also by the dispositions that soldiers bring to the Army and their perceptions of the environmental context encountered after enlistment, regardless of how jobs are clustered. Further, findings suggest that pre-enlistment differences among soldiers in ability and temperament interact with their environmental perceptions in the prediction of various performance outcomes. Considerable variance in soldier performance can be attributed to the main effects of individual differences and environmental perceptions, and generally significant interactions among these factors explain little meaningful variance.

References

Campbell, J., Hanser, L., & Wise, L. (1986, November). The development of a model of Project A criterion space. Paper presented at the 28th Annual Conference of the Military Testing Association, Mystic, Connecticut.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current armed services vocational aptitude battery (ASVAB) composites. (Technical Report No. 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Olson, D. M., & Borman, W. C. (1986). Development and field tests of the Army Work Environment Questionnaire (Working Paper RS-WP-86-06). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

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

Peterson, N., Hough, L., Ashworth, S., & Toquam, J. (1986, November). New predictors of soldier performance. Paper presented at the 28th Annual Conference of the Military Testing Association, Mystic, Connecticut.

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

Staw, B. M., & Ross, J. (1985). Stability in the midst of change: A dispositional approach to job attitudes. Journal of Applied Psychology, 70 (3), 469-480.

NEW PREDICTORS OF SOLDIER PERFORMANCE

Norman Peterson
Leaetta Hough
Steve Ashworth
Jody Toquam

Personnel Decisions Research Institute

Presented on Symposium,

"Project A Concurrent Validation: Preliminary Results"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

NEW PREDICTORS OF SOLDIER PERFORMANCE

Norman Peterson, Leaetta Hough, Steve Ashworth, and Jody Toquam
Personnel Decisions Research Institute

Introduction

New predictors of soldier performance have been developed as part of Project A. Previous papers presented to this association have described the theoretical approach, development, and pilot and field testing of those predictors (Hough, McGue, Kamp, Houston, & Barge, 1985; McHenry & Toquam, 1985; Peterson, 1985; Rosse & Peterson, 1985; Toquam, Dunnette, Corpe, & Houston, 1985). Very briefly, those papers showed that a construct-oriented approach was utilized to identify and develop new measures that would complement the Armed Services Vocational Aptitude Battery (ASVAB) in terms of abilities measured and likelihood of increasing the prediction of training and job performance. Both paper-and-pencil and computer-administered measures were developed to tap constructs in cognitive (primarily spatial) ability, perceptual/psychomotor, temperament, biographical, and vocational interest domains. Pilot and field testing results showed the new measures were psychometrically sound and were measuring constructs relatively unique from the ASVAB.

This paper describes some of the results of analyzing the properties of the new measures, collectively called the Trial Battery, as exhibited in the concurrent validity sample of Project A. This sample consisted of over 9,000 active duty soldiers in their first three years of service, from 19 different military occupational specialties. Other papers in this symposium provide more detailed descriptions of the data collection procedures and job performance criteria also collected from that sample (Harris, 1986; Campbell, Hanser, & Wise, 1986).

New Predictor Factor Scores

The Trial Battery consisted of three major types of instruments: 1) six timed paper-and-pencil tests of cognitive spatial ability, 2) ten computer-administered tests of perceptual/psychomotor ability, and 3) three untimed paper-and-pencil inventories measuring temperament/biographical data (the Assessment of Background and Life Experiences or ABLE), vocational interests (the Army Vocational Interest Inventory or AVOICE), and job reward preferences (the Job Orientation Blank or JOB); collectively referred to as non-cognitive inventories.

Over 60 separate scores are obtained from the full Trial Battery. Space does not allow presentation here of statistics for all these scores. We used principal components factor analysis (varimax rotation) to identify a smaller number of factor scores for use in validity analyses. Examination of these solutions led us to choose 19 factor scores; these were formed by simply summing the scores that defined each factor, not by using a multiple-regression, factor-scoring method. Therefore, we are here using the term factor to denote simply a higher-order organization of Trial Battery test scores, and do not intend these factors as representations of underlying pyschological constructs. These 19 factors are simply a parsimonious method of combining the larger number of individual scale scores for purposes of validity analyses in a way that is faithful to their covariances. Table 1 shows the names of these factors, the number of scores making up the factor, the median reliability coefficients of the scores entering each factor, and the median uniqueness estimate of the factor. Figure 1 shows the names of the scale scores that made up each factor, organized by type of instrument.

The medians of the internal consistency reliability coefficients range

from .46 to .93; mean = .78. All but four are greater than .70. One of these, General Reaction Accuracy, is the sum of percent correct scores on very simple, computerized perceptual tasks. These scores have, by design, severely restricted variance--we were concerned primarily with General Reaction Speed which does have high reliability. The other three factors with relatively low internal consistency reliability are from the Job Orientation Blank, especially the Routine Work and Job Autonomy factors. These are really just single scale scores, with only three or four items on each scale, which probably accounts for the low values.

The test-retest reliabilities range from .13 to .85; mean = .67. The paper-and-pencil measures all have reliabilities of .70 or greater, with the exception of Food Service Interests which is .66. The reliabilities of the computer-administered measures, however, are between .46 and .62, except for the .13 value for General Reaction Accuracy which we discussed above. Although these values are not as high as we would like, keep in mind that these computerized tests are all relatively short (all ten tests are administered in about one hour). Measures that prove most valid could be lengthened to increase reliability. Also, we point out that these are retest intervals of two to four weeks; test-retest coefficients reported for computerized tests are often same-day or next-day intervals which, of course, would yield much higher coefficients.

The uniqueness coefficients in Table 1 are indexes of the amount of reliable variance that does not overlap with, or is unique from, other measures--in this case, the ASVAB. The higher this index, the greater the opportunity for incremental validity (over ASVAB). These values range from .40 to .90; mean = .71. The Trial Battery measures, as a whole, do appear to have high potential for incremental validity, especially for the non-cognitive measures.

In sum, with a few exceptions, the Trial Battery factors appear reliable and relatively unique based on analyses of this large, concurrent validity sample. We add that these results are highly similar to those reported a year ago on a much smaller sample (about 200).

Prediction of Job Performance

Table 2 shows results of initial analyses of the validity of new predictors for predicting job performance and Table 3 shows results of initial analyses of the Trial Battery's incremental validity (over ASVAB) for predicting job performance.

There are five criterion factors shown in both tables. The first two represent "can do" factors and are made up largely of hands-on and written job knowledge test scores (labeled Core Technical Proficiency and General Soldiering Proficiency). The last three represent "will do" factors and are made up largely of peer and supervisor ratings on behaviorally-anchored rating scales and self-reported administrative actions, such as awards and Articles 15 (labeled Effort and Leadership; Personal Discipline; and, Physical Fitness and Military Bearing). As earlier stated, Campbell, et al. (1986) report in more detail the development of these criteria.

Six predictor composites are shown in Table 2, one made up of four factors derived from the ASVAB; the other five made up from the Trial Battery factor scores, combined within instrument type. The composites were formed via multiple regression.

Several things are noteworthy about Table 2. First, it shows the ASVAB does an excellent job of predicting the "can do" criteria, a moderately good job for one of the "will do" factors (Effort), and not very well for two of the "will do" factors. Second, it shows that the Spatial and

Perceptual/Psychomotor composites from the Trial Battery follow a pattern similar to the ASVAB, but do not outpredict the ASVAB. We point out that the perceptual/psychomotor, computer-administered battery requires about 60-75 minutes to administer, but yields validities of .49 and .56 for the "can do" criteria. Also, the six spatial tests require about 90 minutes to administer, and do nearly as well as the ASVAB. Finally, the non-cognitive portions of the Trial Battery do only moderately well at predicting the "can do" criteria, but the ABLE equals or outperforms the ASVAB and the cognitive/perceptual/psychomotor portions of the Trial Battery for predicting the "will do" criteria. Indeed, the ABLE is 13 and 16 points higher than the ASVAB for the Discipline and Fitness/Bearing criteria. All in all, the overall pattern of the findings in Table 2 is about what we expected.

Table 1

Trial Battery Factors, Number of Scores in Each Factor, Median Reliability Coefficients and Uniqueness Estimates of Scores in Each Factor

Composite	Number of Scores	Median Reliability[1] Coefficients: Internal Consistency	Test-Retest	Median Uniqueness[2]
Overall Spatial	6	.85[3]	.70	.55
Psychomotor	6	.80	.62	.71
Perceptual Speed/Accuracy	6	.80	.57	.72
Number Speed/Accuracy	4	.91	.58	.67
General Reaction Speed	2	.93	.46	.90
General Reaction Accuracy	2	.52	.13	.45
Achievement	3	.82	.78	.81
Dependability	2	.77	.77	.74
Adjustment	1	.81	.74	.79
Physical Condition	1	.84	.85	.83
Skilled Technician Interests	7	.89	.75	.82
Structure/Machines Interests	4	.92	.81	.75
Combat-Related Interests	3	.90	.80	.75
Audiovisual Arts Interests	3	.83	.74	.81
Food Service Interests	2	.81	.66	.78
Protective Service Interests	2	.83	.76	.81
Organization/Co-Worker Support	4	.67	N/A	.65
Routine Work	1	.46	N/A	.40
Job Autonomy	1	.50	N/A	.47

Note: N varies, but all > 7,000

[1] These are odd-even coefficients, corrected with Spearman Brown procedure, or coefficient Alpha for internal consistency and correlations over a two-four week interval, N=470, for test-retest.

[2] Uniqueness = $R - R^2$, where R = internal consistency reliability estimate and R^2 = squared multiple correlation of all ASVAB tests with each new predictor.

[3] This is based on a separately-timed, split-half coefficient collected during pilot testing, N = 118, because some of these tests are speeded, making odd-even coefficients inappropriate.

FROM PAPER-AND-PENCIL TESTS

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

FROM COMPUTERIZED MEASURES

Psychomotor
- Cannon Shoot Test (Time Score)
- Target Shoot Test (Time To Fire)
- Target Shoot Test (Log Distance)
- Target Tracking 1 (Log Distance)
- Target Tracking 2 (Log Distance)
- Pooled Mean Movement Time

Perceptual Speed and Accuracy
- Short Term Memory Test (Percent Correct)
- Perceptual Speed & Accuracy Test (Decision Time)
- Perceptual Speed & Accuracy Test (Percent Correct)
- Target Identification Test (Decision Time)
- Target Identification Test (Percent Correct)

Number Speed and Accuracy
- Number Memory Test (Percent Correct)
- Number Memory Test (Initial Decision Time)
- Number Memory Test (Mean Operations Decision Time)
- Number Memory Test (Final Decision Time)

General Reaction Speed
- Choice Reaction Time
- Simple Reaction Time

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

FROM NON-COGNITIVE INVENTORIES

Organizational and Co-Worker Support (JOB)
- Job Pride
- Job Security Comfort
- Serving Others
- Ambition

Routine Work (JOB)
- Routine

FROM NON-COGNITIVE (CONTINUED):

Job Autonomy (JOB)
- Autonomy

Achievement (ABLE)
- Self-Esteem Scale
- Work Orientation Scale
- Energy Level Scale

Dependability (ABLE)
- Conscientiousness Scale
- Non-Delinquency Scale

Adjustment (ABLE)
- Emotional Stability Scale

Physical Condition (ABLE)
- Physical Condition Scale

Skilled Technician Interest (AVOICE)
- Clerical/Administrative
- Medical Services
- Leadership/Guidance
- Science/Chemical
- Data Processing
- Mathematics
- Electronic Communications

Structural/Machines Interest (AVOICE)
- Mechanics
- Heavy Construction
- Electronics
- Vehicle/Equipment Operator

Combat Related Interest (AVOICE)
- Combat
- Rugged Individualism
- Firearms Enthusiast

Audiovisual Arts Interest (AVOICE)
- Drafting
- Audiographics
- Aesthetics

Food Service Interest (AVOICE)
- Food Service Professional
- Food Service Employee

Protective Services Interest (AVOICE)
- Law Enforcement
- Fire Protection

Figure 1. Test and inventory scale scores making up Trial Battery Predictor Factors.

Table 2

Multiple Correlation[1] of Six Independent Predictor Composites with Each of Five Job Performance Criterion Factors.

	PREDICTORS					
CRITERION FACTORS	ASVAB[2] Composite K = 4	Spatial Abilities Composite K = 1	Perceptual/ Psychomotor Abilities Composite (Computerized) K = 5	JOB Composite (Preferences) K = 3	ABLE Composite (Temperament/ Biodata) K = 4	AVOICE Composite (Interests) K = 6
1. Core Technical Proficiency	.60	.54	.49	.26	.24	.33
2. General Soldiering Proficiency	.66	.64	.56	.29	.25	.37
3. Effort and Leadership	.35	.28	.27	.19	.34	.26
4. Personal Discipline	.19	.16	.14	.11	.32	.15
5. Physical Fitness & Military Bearing	.21	.11	.11	.12	.37	.12

Note: Entries in the table are averaged across 9 Army MOS with complete sets of Job Performance Criterion measures. Total sample size is 3902. Sample sizes range from 281 to 570; median = 432.

[1] Multiple Rs are adjusted for shrinkage and corrected for restriction in range, but not corrected for criterion unreliability.

[2] K = the number of predictor scores in the composite.

Table 3

Increments in Multiple Correlations[1] (Over R Using ASVAB Composite) as A Function of Adding Trial Battery Factor Scores for Each of Five Job Performance Criterion Factors.

	CRITERION FACTORS				
PREDICTOR	Core Technical Proficiency	General Soldiering Proficiency	Effort and Leadership	Personal Discipline	Fitness & Bearing
ASVAB[2] Composite Alone (K = 4)	.60	.66	.35	.19	.20
ASVAB Plus Trial Battery Factors (K = 23)	.64	.70	.45	.37	.42
Increment	.04	.04	.10	.18	.22

Note: Entries in the table are averaged over 9 Army MOS with complete sets of criterion measures. Total sample size is 3902. Sample sizes within MOS range from 281 to 570; median = 432.

[1] Multiple Rs are adjusted for shrinkage and corrected for restriction in range, but not corrected for criterion unreliability.

[2] K = the number of predictor scores in the composite.

While the AVOICE does not show higher prediction than the ASVAB for the "can do" criteria, it is interesting that it correlates .33 and .37 with those criteria. The AVOICE was intended primarily to assist in classification rather than prediction per se, so it is encouraging to see these correlations with "can do" criteria. Finally, with respect to Table 2, we note that the JOB, ABLE, and AVOICE are expected to add most to the prediction of attrition; those analyses have not been done yet.

Table 3 shows a first, very crude look at the incremental validity of the Trial Battery. In these analyses, we simply added all 19 Trial Battery Factor scores to the ASVAB factor scores and looked at the increase in the mulitiple correlation. The third row in Table 3 shows that 1) the prediction of all five criteria is increased, 2) little increase occurs for the "can do" criteria, and 3) sizeable increases occur for the "will do" criteria.

Efforts are underway now to make more refined Trial Battery composites and to estimate the classification efficiency increments obtained via use of the Trial Battery. These initial results, however, show that the new predictors do 1) predict soldiers' job performance at meaningful levels in the way that was expected and 2) make meaningful increments over the ASVAB to validity for important aspects of soldiers' job performance.

References

Campbell, J., Hanser, L., & Wise, L. (1986). *The development of a model of Project A criterion space.* Paper presented at the 28th Annual Military Testing Association Conference Mystic, Connecticut.

Harris, J. (1986). *The Project A concurrent validation data collection.* Paper presented at the 28th Annual Military Testing Association Conference, Mystic, Connecticut.

Hough, L. M., McGue, M. K., Kamp, J. D., Houston, J. S., & Barge, B. N. (1985). *Measuring personal attributes: Temperament, biodata, and interests.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

McHenry, J., & Toquam, J. L. (1985). *Computerized assessment of perceptual and psychomotor abilities.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

Peterson, N. G. (1985). *Mapping predictors to criterion space: Overview.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

Rosse, R. L., & Peterson, N. G. (1985). *Using microcomputers for assessment: Practical problems and solutions.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

Toquam, J. L., Dunnette, M. D., Corpe, V. A., & Houston, J. S. (1985). *Adding to the ASVAB: Cognitive paper-and-pencil measures.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

Note: This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences. Contract Number MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

EFFECT OF PRACTICE ON SOLDIER TASK PERFORMANCE

Paul Radtke
Dorothy S. Edwards

American Institutes for Research

Presented on Symposium,

"Job Performance: What Do Soldiers Know, What Can They Do?"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Effect of Practice on Soldier Task Performance*

Paul Radtke
Dorothy S. Edwards

American Institutes for Research

One of the forms administered in the Army's Selection and Classification study, usually known as Project A, was a Job History Questionnaire. For each of nine Military Occupational Specialties (MOSs) the form listed all of the tasks covered by paper and pencil knowledge tests and by hands-on performance tests. These tasks were selected from the domain of tasks for an MOS by a panel of experts because they were done frequently and were important to overall job performance. About thirty tasks were selected for each MOS; all were measured with performance based knowledge tests; about half were also measured with hands-on tests.

In the Job History Questionnaire soldiers were asked to indicate how often during the past six months they had performed each task, using a scale of "Not at all, 1-2 times, 3-5 times, 6-10 times, or more than 10 times." Next, soldiers indicated how recently they had performed each task, using a scale of "Never, during past month, 1-3 months ago, 4-6 months ago, or more than 6 months ago."

The frequency and recency ratings were correlated with the scores on the knowledge tests and with the hands-on tests for each MOS. The results for two sample MOSs, one combat and one support MOS, are shown in Tables 1-2. The number of cases for these correlations varies, but in every case is substantial. The minimum and maximum N is given at the top to reduce the number of columns in the tables. When there is a wide range in the number of cases it reflects a smaller N on one or two tests and nearly maximum Ns on the others. The size of the N makes a rather small correlation significant statistically; the rather small correlations probably have little practical significance. Note that the recency correlations should be negative, because of the way the scale was written.

The tables have some items of interest, however. Recency appears to be more closely associated with test performance than does frequency of practice, in that more of these correlations attain statistical significance. Recency and frequency are correlated, as shown in the last column of the tables.

There is a tendency for the more complex tasks to be more highly correlated with frequency and recency, though there are some exceptions in both directions -- complex tasks not correlated or easy tasks correlated.

Performance on MOS-specific tasks tends to be more highly correlated with frequency and recency of practice than performance on the common

---

*This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Table 1. Correlation between Job History Questionnaire Scales (Frequency and Recency of Performance) and Task Test Scores for 11B (*Infantryman*)

Decimal points omitted; * = significant at P = .01

	K Tests (N=495-697)		NO Tests (N=496-696)		Frequency & Recency
11 B Knowledge	Freq.	Rec.	Freq.	Rec.	
Perform CPR	04	-17*			-61
Adm Nerve Agent Antidote	04	-13*			-48
Put on Field/Pres. Dressing	03	-06	09	-09	-41
Perform OP Maint. on M16A1	06	-08	08	-04	-40
Load/Reduce/Clear M60	12*	-09	25*	-20*	-50
Engage w Hand Grenades	08	-13*	05	-08	-46
Prepare Dragon for Firing	14*	-16*	18*	-18*	-57
Prepare Range Card for M60	00	-10	24*	-19*	-54
Call for/Adjust Indirect Fire	16*	-16*			-60
Navigate on the Ground	16*	-20*			-52
Id Terrain Features on MAP	10*	-10*			-37
Put on M17 Mask	02	-05	07	-08	-48
Put on Protective Clothing	02	-13*			-39
Collect/Report Info	-06	-06			-38
Camouflage Self/Equip	07	-07			-42
Id Armored Vehicles	15*	-13*			-45
Move under Direct Fire	00	-03			-63
Estimate Range	-02	04			-58
Move over Obstacles	04	-01			-52
Operate Radio Set AN/PRC-77	09	-10*	04	-02	-50
Install/Fire Claymore Mine	07	-07	22*	-19*	-45
Tech of Urban Terr Movement	-06	03	-03	-04	-49
Select Nasty Urban Firing Pos	-02	-03			-64
Establish Obv Post	01	-08			-64
Sel Fire Team/Overwatch Pos	02	-03			-72
Zero AN/PVS-4 to M16A1	-02	-02	-06	-17*	-65
Place AN/PVS-5 into operation	09	-04			-60
Set Headspace/Timing on .50	37*	-37*	41*	-35*	-72
Engage Target w LAW			-04	-03	-52

Table 2
71L (Administrative Specialist)

	K Tests (N=498-508)		NO Tests (N=494-508)		Frequency & Recency
71L Knowledge	Freq.	Rec.	Freq.	Rec.	
Adm Nerve Agent Anti-Self	07	-20*			-49
Load/Clear M16A1	-01	-01			-32
Oper Maint M16A1	06	-06	02	-10	-36
Det Magnetic Azimuth	-02	-08			-34
Det Grid Coordinates	08	-17*	09	-16*	-40
Put on M17 Mask	-05	03	07	-21*	-34
Maintain M17 Mask	04	-09			-46
Put on Protective Clothing	18*	-18*			-45
Know Rights as POW	09	-15*			-63
Camouflage Self/Equip	14*	-12*			-55
Prac Noise/Light/Litter Disc	-16*	-16*			-67
File Documents/Corresp	18*	-13*	10	-12*	-80
Est Functional Files	04	-06			-72
Control Supplies	07	-09			-88
Rec/Contl Office Equip	02	00			-84
Dispatch Outgoing Dist.	11*	-10*			-79
Type Military Orders	01	-02	10	-11*	-77
Type 2nd Comment to DF	32*	-30*	12*	-11*	-80
Type Jt Message Form	15*	-18*	08	-14*	-79
Type a Memo	14*	-19*	08	-14*	-76
Type a Basic Comment to DF	20*	-20*	20*	-24*	-80
Assemble Correspondence	16*	-13*			-79
Type Military Letter	25*	-24*	18*	-15*	-80
Safeguard FOUO Material	03	-04			-84
Rec/Trans Classified Material	10	-06	16*	-12*	-82
Put on Field/Press Dressing			-02	-13*	-34
Prep. Requisition/AUTODIN			17*	-14*	-75

tasks. It may be that common tasks have been subject to more practice during the soldier's enlistment. This hypothesis is consistent with the generally higher mean scores on the common tasks. If true, the common tasks may have been "overlearned," and thus less subject to forgetting or to decrement through lack of practice.

Some common tasks were tested in more than one MOS. This allows us another way to look for consistency in association of test scores and frequency or recency of practice. Table 3 shows these data for the common tasks. One task, "Determine grid coordinates" shows significant correlations with frequency and recency in six of the seven MOSs in which the knowledge test was given. It also showed significant correlations in the hands-on tests in most of the MOSs in which it was given. It is the consistency of the findings rather than the magnitude of the actual correlations that makes us believe that competency in this task is indeed related to frequency and recency of practice. The test was very similar in both measurement methods: soldiers had to read grid coordinates using a protractor. They had an advantage in the written mode in that the correct answer appeared as one of four choices, whereas they had to report the coordinates to the test administrator in the hands-on mode without the recognition advantage afforded by the multiple choice item.

A second test that has a similar pattern of significant correlations with the knowledge tests is "Put on and wear protective clothing." This test, however, does not correlate with the hands-on measure. Since the soldier must put on the clothing required at four progressive levels of protection, over-dressing at phase 1, or MOPP Level 1, as it is called, could keep the soldier from correctly reaching the higher levels.

Naturally we looked for characteristics that these two tasks have in common that are not present in other tasks that do not show this pattern of correlations. We found only one. Each of the tasks requires a specific procedure that terminates in an objectively verifiable product or result. Exact grid coordinates are determined and reported, and certain garments are worn at each MOPP level. This means that the "right answers" are totally unequivocal and readily observable by even a careless scorer in the hands-on mode. These tests had reliability estimates that were among the highest in the MOSs in which they appeared, which is probably also a function of the clarity and observability of the response.

Another test that is fairly consistent in correlations with frequency and recency is "Load, reduce, and clear the M60 machinegun." It was given in only three MOSs, so the consistency cannot be as pronounced as with the grid coordinates and protective clothing tests. Table 4 shows the correlations for this task as well as those for a similar task: "Load, reduce, and clear the M16A1 rifle." Performance on the M16 tests is not as highly correlated with frequency, probably because it is the soldier's main weapon and is more often practiced and proficiency is maintained at a high level. The task is also somewhat simpler than the matching task on the M60.

At the bottom of Table 4 we have shown the mean percent passing the knowledge tests and the mean percent "GO" on the hands-on test for all MOSs in which the M60 and M16 tasks were covered. Note that performance on the hands-on test is higher than on the knowledge test for both tasks,

Table 3. Correlations Between Job History Questionnaire Scales and Scores on Common Soldiering Tasks
Decimal points omitted: *= significant at P = .01

A. Frequency - Knowledge Tests

K Tests	11B	13B	19E	31C	63B	64C	71L	91A	95B
CPR	04	-02		03		09		00	-11*
Nerve agent	04	02	01			07	07		
F/P dressing	03		10	15*	09	07		-03	10*
LRC M16		-09		-01	01	-02	-01	07	00
Op/Mtn M16	06			-03	05	03	06		
LRC M60	12*					18*			20*
Mag. Azim.					04		-02		12*
Grid Coor.			15*	16*	13*	13*	08	18*	08
Put on mask	02	07			05	05	-05		08
MOPP	02	07	12*	09	16*	15*	18*	14*	
CEOI			38*						24*

B. Recency - Knowledge Tests

CPR	-17*	02		-09		-10*		00	-13*
Nerve agent	-13*	-12*	-11*			-06	-20*		
F/P dressing	-06		-04	-12*	-07	-08		-05	-08
LRC M16		00		-12*	-02	-01	01	-16*	-02
Op/Mtn M16	-08			00	-02	-10*	-06		
LRC M60	-09					-15*			-22*
Mag. Azim.					-06		-08		-05
Grid Coor.			-12*	-23*	-18*	-18*	-17*	-20*	-04
Put on mask	-05	-04			00	-04	03		-06
MOPP	-13*	-04	-12*	-15*	-11*	-12*	-18*	-12*	
CEOI			-29*						-27*

C. Frequency - HO Tests

CPR		12*				15*		13*	17*
Nerve agent		01				11*			
F/P dressing	09		06	04	02	01	-02	07	14*
LRC M16		01		-03	-01	01			06
Op/Mtn M16	08					05	02		
LRC M60	25*					14*			07
Mag. Azim.					01				07
Grid Coor.			09	22*		13*	09	18*	17*
Put on mask	07	04			-04	08	07		09
MOPP		05		01		04			
CEOI			31*						

D. Recency - HO Tests

CPR		-19*				-18*		-09	-15*
Nerve agent		-01				-10*			
F/P dressing	-09		-10	-15*	-06	-08	-02	-11	-13*
LRC M16		01		-07	-02	02			-12*
Op/Mtn M16	-04					-07	-10		
LRC M60	-20*					-30*			-03
Mag. Azim.					-12*				-02
Grid Coor.			-06	-21*		-17*	-16*	-19*	-13*
Put on mask	-08	00			-02	-06	-21*		-06
MOPP		-06		-07		-08			
CEOI			-27*						

but the performance on the M16 weapon is superior to performance on the M60. The M60 task is somewhat more complex, and has more steps, but the M16 is almost certainly practiced more often. Soldiers do appear to be able to load, reduce, and clear their primary weapon, as indicated by the mean of 85% GO on the hands-on test.

Table 4. Correlations between frequency and recency of practice and test scores on two weapons, the M60 machinegun and the M16 rifle

	Infantry	Cannon Crewman	Radio Operator	Auto Mechanic	Truck Driver	Admin. Specialist	Medic	Military Police
Frequency								
LRC M60 K	12*				18*			20*
LRC M60 HO	25*				14*			07
LRC M16 K		-09	-01	01	-02	-01	07	00
LRC M16 HO		01	-03	-01	01			06
Recency								
LRC M60 K	-09				-15*			-22*
LRC M60 HO	-20*				-30*			-03
LRC M16 K		00	-12	-02	-01	01	-16*	-02
LRC M16 HO		01	-07	-02	02			-12*

	LRC M16	LRC M60
Mean % correct, K	72.79	61.80
Mean % GO, H-O	85.84	68.35

A final test that shows substantial correlation with both frequency and recency of practice is "Use automated CEOI" (Communications Electronics Operating Instructions). It was given in only two MOSs, and is similar to

grid coordinates in that it results in an objectively observable result. The correlations were as shown below:

	Tank Crewman	MP
CEOI K-test & freq.	38*	24*
CEOI K-test & recency	-29*	-27*
CEOI H-O test & freq.	31*	Not given
CEOI H-O test & recency	-27*	Not given

This test requires memory of procedures for looking up information in a table and reporting call signs, radio frequencies, and authentication data. A number of soldiers taking the hands-on test reported on how easily the procedures for reading the table are forgotten.

Conclusions

The ratings on frequency and recency of practice of tasks tested in Project A show very low correlations with test performance. There are, however, some tasks that show a significant relationship, and in a consistent enough manner to suggest that we are not dealing with chance results.

Tasks that are related to practice seem to be those that produce objectively observable results, that are relatively complex, and related to the MOS specific parts of the job rather than to the common soldier tasks.

Reference

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

Number and percent of common and MOS-specific task tests that are significantly related to frequency and recency of practice

FREQUENCY

	Knowledge Tests			Hands-on Tests		
	No. tests	No. sig.	%	No. tests	No. sig.	%
Common	120	30	25	60	20	33
MOS-specific	143	40	28	75	28	37
Total	263	70	27	135	48	35

RECENCY

	Knowledge Tests			Hands-on Tests		
	No. tests	No. sig.	%	No. tests	No. sig.	%
Common	120	44	37	60	26	43
MOS-specific	143	42	29	75	38	51
Total	263	86	33	135	64	47

SOME CONDITIONS AFFECTING ASSESSMENT OF JOB REQUIREMENTS

Elizabeth P. Smith
Paul G. Rossmeissl

U.S. Army Research Institute

Presented on Session, "Improving Training Performance"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Some Conditions Affecting Assessment of Job Requirements

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

Paul G. Rossmeissl[2]
Hay Systems, Inc.

As an adjunct to the Army Research Institute's Project A to improve the selection and classification process, research was initiated to develop and test a rating scale method to assess (Eaton, et. al., 1984) human attributes (e.g., abilities, interests, etc.) that are needed for success in a particular Military Occupational Specialty (MOS) (Smith, 1985). The work followed from the ability taxonomy and rating scale work by Fleishman and his associates (see Fleishman & Quaintance, 1984). Within Project A, a taxonomy of human attributes that affect performance was developed from expert judgments of validity (Wing, Peterson, & Hoffman, 1984). The taxonomy included 21 clusters of cognitive/perceptual, psychomotor, and noncognitive (temperament and interests) variables. Smith (1985) constructed a set of scales corresponding to 20 of these attributes plus physical strength and stamina. This set of scales, the Attribute Assessment Scale (AAS), which was designed to use work supervisors as Subject Matter Experts (SMEs), contains primarily Army-specific behavioral anchors. Several problems were uncovered during preliminary tests of the instrument with two different samples (Smith & Rossmeissl, in process). The research which is presented here attempted to address those issues. As with the earlier research, the goal was to demonstrate that the scales can produce reliable, differential profiles of attribute requirements that discriminate across MOS. These profiles then could be matched to measures of an individual's attributes for selection and classification purposes.

In the first test of the AAS (Smith, 1985), senior noncommissioned officers (NCOs) from two MOS provided ratings of the requirements for entry level work in their own MOS for three performance levels (15th, 50th, and 85th percentiles). Two types of Intraclass Correlation Coefficients (ICCs) were calculated over all attributes. The first (r_1) provides a point estimate of interrater reliability or the reliability of a single rater. The second (r_k) indicates the reliability of the mean rating. These coefficients were extremely weak. There was very little interrater agreement and at least 30 raters were needed to obtain moderately reliable means--a number higher than would be practical in operational use. An ANOVA indicated that attribute profiles for the two MOS were not significantly different.

There appeared to be three major problems related to the instrument and the research. First, the inclusion of three performance levels may have had a strong, negative impact on the results. The demands of the task appeared

---

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

[2]Affiliated with U.S. Army Research Institute at the time this research took place.

to impose a unique kind of restriction in the range of possible ratings, plus it took considerable effort. Second, the multiple levels added more confusion to a performance criterion which was already very broad -- all work within all duty positions -- allowing for considerable variance. The third problem centered on the scale anchors. This included SMEs' frustration with their content and/or difficulty in using them as reference points for evaluating duties within their MOS.

The second test of the AAS (Smith & Rossmeissl, in process) considered two of these issues. SMEs were a small number of officers and NCOs from three MOS. We provided a written job description from Army Regulation 611-201, and SMEs gave a single rating of the level of each attribute required for "average" performance of entry level work in their own MOS. An important aspect of the research was a post-rating discussion period during which SMEs provided information about problems that they had in completing the task, specific issues related to interpretation of "average" performance, confidence in their responses, and ways to improve the procedures.

With the exception of one MOS for which procedural problems were noted, the results were promising. Overall, the magnitudes of the ICCs were better than those obtained in the original research. Reliabilities of mean rating (r_k) equal to .73 and .84 with only 4 and 9 raters respectively were encouraging. ANOVA results indicated no significant differences in profiles across MOS, but given the small sample sizes this was not surprising. Our post-rating discussions indicated that use of the criterion "average" performance may have reduced MOS differences as well. Problems with this terminology included some tendency a) to describe the average soldier rather than, e.g., the average Administrative Specialist, b) to confuse average performance with average level of requirements, and c) to view average performance as actually substandard. The discussions also confirmed there were still problems related to the anchors and the ambiguity/enormity of the "whole job" criterion.

Given these outcomes, we decided to test the rating scales again under different conditions. In this research we examined ratings of attribute requirements for the whole MOS versus ratings of important, representative component tasks using two sets of scales with different anchors.

METHOD

Sample

One hundred fifty-nine NCOs from three MOS (Cannon Crewman: 13B, Light Wheel Vehicle Mechanic: 63B, and Single Channel Radio Operator: 31C) at two posts served as SMEs.

Procedure

Within MOS and posts, SMEs were assigned in blocks of 12 or less to one of 4 condition groups. Group I rated the job as a whole, using the original, behaviorally-anchored AAS. Group II rated the job as a whole, using scales with generic anchors (1=very low, 4=moderate, 7=very high). Groups III and IV rated the attribute requirements for 15 component tasks of their MOS. The tasks were those used in the hands-on testing portion of Project A. Group III used the behaviorally-anchored scales; Group IV, the generically-anchored ones. SMEs estimated the levels of the 22 attributes

which are required for "successful performance" of Skill Level 1 work for their own MOS. SMEs in the previous research favored this choice of performance criterion. In addition to the written instructions, we provided SMEs with brief training in how to use the scales to derive ratings.

Analyses

To determine reliability, we calculated ICCs (r_1 and r_k) from Attribute X Rater ANOVAs by group for each MOS. To compare reliabilities based on same sized groups, we estimated reliability of mean ratings based on 6 raters (r_6) using the Spearman-Brown formula. We performed an MOS X Attributes X Anchor (Generic vs. Behavioral) X Criterion (Whole Job vs. Tasks) univariate repeated-measures ANOVA to examine differences in profiles among MOS and any effects due to anchor or criterion conditions. The single, highest rating assigned to any task within each attribute was used in the ANOVA.

Results

The ICCs (r_1, r_k, r_6) for the four conditions by MOS are given in Table 1. Overall, estimates of interrater agreement are low. The best r_1s are for Radio Operators across all 4 conditions, yet there still are large between-subjects variances for all MOS. Across MOS, no particular condition yielded higher r_1s or r_ks than another.

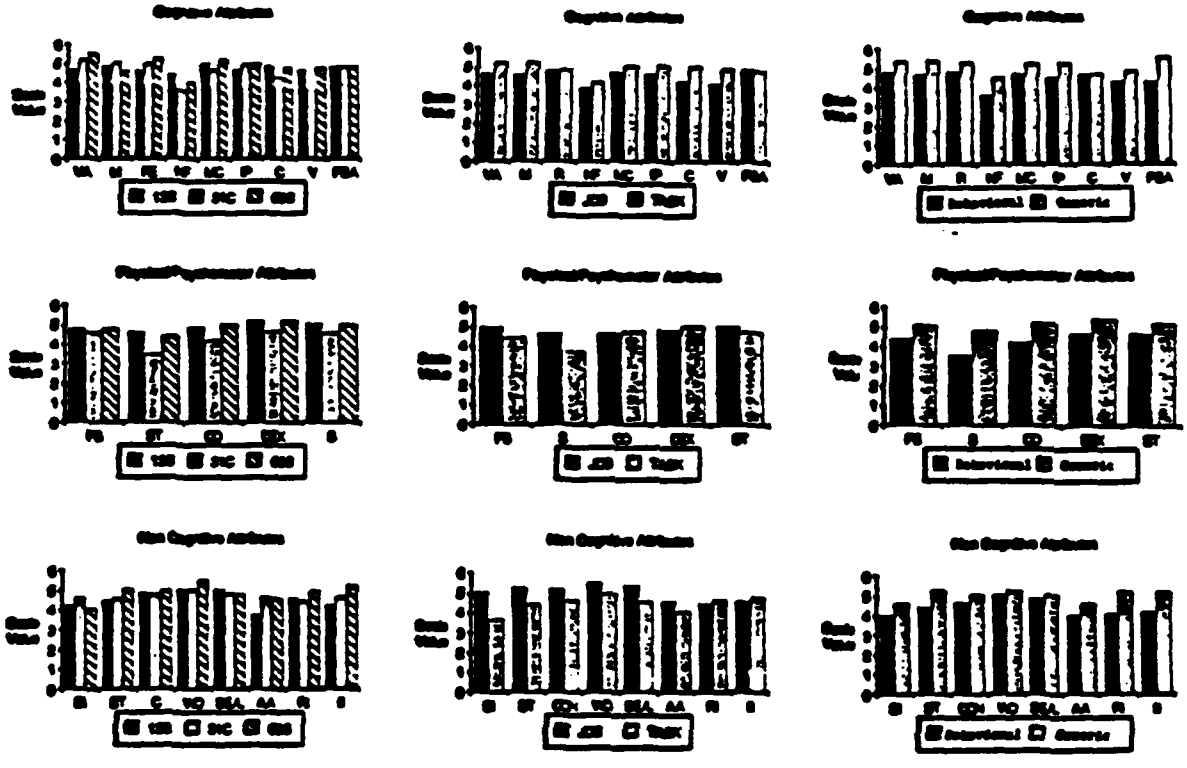
Table 1

Reliability estimates for a single rater, mean of k raters, and mean of six raters of three MOS by experimental conditions

MOS	Anchor Type	Criterion	k	r_1	r_k	r_6
Cannon Crewman	Behavioral	Task	19	.08	.63	.34
		Job	12	.22	.77	.63
	Generic	Task	25	.07	.67	.31
		Job	12	.04	.36	.20
Radio Operator	Behavioral	Task	9	.28	.77	.70
		Job	12	.22	.77	.63
	Generic	Task	12	.17	.71	.55
		Job	17	.19	.80	.58
Mechanic	Behavioral	Task	22	.11	.74	.43
		Job	7	.12	.48	.45
	Generic	Task	6	.07	.32	.32
		Job	6	.21	.61	.61

The MOS X Attribute X Anchor X Criterion ANOVA indicated there are significant differences in attribute profiles across MOS and that these differences were affected by the experimental conditions. Although the 4-way interaction is not significant, two 3-way interactions (Attribute X MOS X Anchor and Attribute X MOS X Criterion) and all 2-way interactions involving attribute are significant with a Geisser-Greenhouse $p<.05$. That is, mean

a function of the type of anchors or the criterion. Collapsing over type of criterion, generically-anchored scales yielded higher mean ratings for all attributes. On the other hand, the effects of criterion condition (job vs. tasks) were dependent on the type of attribute. For the most part, across MOS, we found higher means for evaluations of the whole job for the cognitive/perceptual attributes, some of the psychomotor attributes, and of the noncognitive attributes, realistic and investigative interests. The opposite was true for physical strength, stamina, and the other noncognitive (temperament) attributes. Figures 1(a-c) graphically depict the three 2-way interactions.



a b c

Attributes:

Cognitive

VA • Verbal Ability
M • Memory
R • Reasoning
NF • Number Facility
MC • Mechanical Comprehension
IP • Information Processing
C • Closure
V • Visualization
PSA • Perceptual Speed & Accuracy

Physical/Psychomotor

PS • Physical Strength
S • Stamina
CO • Coordination
DEX • Dexterity
ST • Steadiness

Non-Cognitive

SI • Social Interaction
ST • Stress Tolerance
CON • Conscientiousness
WO • Work Orientation
SE/L • Self-Esteem Leadership
AA • Athletic Ability
RI • Realistic Interests
II • Investigative Interests

Figure 1. Comparigon of profiles of attribute means by a MOS, b Criterion (Job vs. Task), and c Anchor (Behavioral vs. Generic).

DISCUSSION

As with initial tests of the AAS, the interrater agreement found here is relatively low. For most purposes, however, we are more interested in the reliability of the mean ratings which were moderate for most of the conditions. Use of generically-anchored scales did not improve the reliabilities as our previous research had suggested, but the behaviorally anchored scales were no more reliable than the generic anchors. In effect, however, using behavioral anchors tended to lower mean ratings, perhaps by reducing a "more means better" tendency toward inflating estimates of requirements for good performance. These findings suggest that in similar situations the impact of using behavioral based anchors may not merit their increased developmental effort and cost.

Similarly, to the degree it was tested here, having SMEs rate components of the job did not increase agreement among raters either. In our analyses we used only one of the 15 ratings made by SMEs in the task rating conditions. Perhaps we would find better interrater reliability if we focused on each task individually. The choice of criterion did affect magnitude of ratings, but not in the same way for all attributes. Differences in means, as well as lack of agreement among raters, may well have been a function of the comprehensiveness or representativeness of the tasks. Some SMEs argued that the specific tasks we used required little or none of some attributes (especially temperament attributes), but that these attributes are required for other aspects of the job. A few SMEs indicated they gave high ratings on the tasks for this reason, thus ignoring our instructions to rate only the 15 tasks provided.

Although we were unable to increase reliability by altering the conditions of the administration of the AAS, the data were sufficiently reliable to yield meaningful results. The key interaction of MOS and attribute was statistically significant: We did attain significantly different requirements profiles across MOS mean ratings. Also significant were the comparisons investigating the effects of anchor type and level of analysis (job versus task). In other words, while the reliabilities were low, they were sufficient to provide valuable information. Given this and the other findings, the AAS, while not producing results which advocate its use for selection and classification purposes, still may have some potential. For example, it may be useful for identification of the two-three top high-driver attributes for an MOS, or for evaluation of a narrowly defined task, such as a particular kind of mission. Our debriefings with SMEs lead us to believe that any future use of the AAS or similar kinds of instruments really should involve an intensive training session. SMEs should be given thorough explanations, with examples, of what the attributes entail and helped to see how they relate to various aspects of the job.

REFERENCES

Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (October, 1984). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1984 Fiscal Year. (Technical Report No. 660). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Fleishman, E. A., & Quaintance, M. K. (91984). Taxonomies of human performance. Orlando, FL: Academic Press Inc.

Smith, E. P. (1985, November). Developing new attribute requirements scales for military jobs. Proceedings of the 27th Annual Conference of the Military Testing Association, San Diego, CA.

Smith, E. P. & Rossmeissl, P.G. (1984). Attribute assessment: Initial test of scales for determining human requirements of military jobs. Technical Report. U.S. Army Research Institute for the Behavioral Sciences, Alexandria, VA.

Wing, H., Peterson, N. G., & Hoffman, R. G. (1984, August). Expert judgments of predictor-criterion validity relationships. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Canada.

SHORT VERSUS LONG TERM TENURE AS A CRITERION FOR VALIDATING BIODATA

Elizabeth P. Smith
Clinton B. Walker

U.S. Army Research Institute

Presented on Session, "Validating Selection Criteria"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

Short Versus Long Term Tenure as a Criterion for Validating Biodata

Elizabeth P. Smith and Clinton B. Walker[1]
U. S. Army Research Institute for the Behavioral and Social Sciences

This research tests the hypothesis that the traditional criterion for validating biodata in military research, viz. attrition during the first six months of service versus successful completion of that period, has produced less effective scoring keys and lower validities than a longer criterion period would. This hypothesis is based on two findings. First, at least half of attritions in previous research have occured after the first six months of service (Goodstadt & Yedlin, 1980; Hicks, 1981). Second, only half as many items in a 60-item biodata instrument were keyable at the six-month point as were at tenures of one to three years in data from 5,941 applicants to the Army in FY1981 and 1982 (Walker, 1985). If these findings are generally true, then keying on tenures longer than six months will move many first term attritions from the successful criterion group to the unsuccessful one, where they belong, and will produce a larger pool of keyable items. Both of those results should improve validity. In the present paper, items from the Army's Military Applicant Profile (MAP) are keyed on status at the 6-month and then at the 39 - 45 month point, depending on date of entry, and the validities are compared for those two criterion periods.

Method

Instrument

A 240-question research version of the MAP, which is a multiple choice biodata questionnaire, provided the items. Two forms of the instrument, with different sequences of the items, were used. In content, the questions deal with self-esteem, motives for enlisting, experiences in school, work experience, expectations of military life, social habits, experiences in the family, athletic activity, and miscellaneous other experiences.

Sample

The sample was 9,416 receptees at all seven Army Reception Stations who took the instrument in January-June 1982. This number included 7,653 males, of which 6,403 were high school graduates and 1,250 were non-graduates or GED holders. Also in the sample, but examined only for cross-validity, were 1,763 females, all high school graduates.

Criteria

All cases were divided into "stayers" and "leavers" as follows. Stayers were either on Active Duty at the end of the period being examined or had

---

[1]The opinions in this paper are the authors' and do not necessarily reflect views or policy of the Army Research Institute or the Department of the Army. Richardson, Bellows, Henry, and Company, Inc., under contract to Army Research Institute, developed the items for this work and collected the raw predictor data and six-month criterion data. Joseph Stephenson created the dataset with the longer tenures. We gratefully acknowledge his support.

been discharged for positive reasons (e.g., end of enlistment, transfer into an officer candidate program) or "no fault" reasons (e.g. medical, hardship). Leavers were cases who had been discharged for any causes other than those above. These latter cases were presumed to have been discharged early for any of various failures to adapt to Army life. For comparing short and longer tenures as criteria, the status of the cases was examined first at the end of the initial six months of service and then as of 1 October 1985, which was from 39 to 45 months after accession. Leavers after the first six months were in the successful group for the first analysis and in the unsuccessful group for the longer tenure.

Procedure

Empirically derived scoring keys were developed on a 60% sample of all of the males. To select items for keying, we ran item-level chi square tests on the frequencies with which the separate response choices were picked by the criterion groups (stay vs. leave). Items giving $p < .05$ were retained for keying. These items were keyed using a horizontal percentage method (Cascio, 1982; Riegelhaupt & Bonczar, 1985), weighted for differences in sizes of the criterion groups. These weighted percentages of stayers were then rounded and converted to single digit weights ranging from -1 to +3. The conversion rule was as follows: up to 24% stayers = -1; 25 to 34% = 0; 35 to 44% = 1; 45 to 54% = 2; >54% = 3. Under this rule for assigning weights, some items were weighted more heavily than others by having a wider range of possible scores.

Item scores for each case were summed and tested for differences between criterion groups. Then, point biserials were calculated on the relation between total scores and the dichotomous stay-leave criterion. After finding validities on the development sample, we computed validities on the independent holdout sample of all males, on two random samples of the females 60% and 40%), and on similar splits of the two male groups (graduates and on-graduates) which were subsets of the larger development and holdout groups. These procedures were followed first for the short criterion period maximum service of six months) and then, on the same cases, for the longer criterion period.

As a check on whether the same items would be effective for predicting success over both short and long criterion periods, we divided items into those which were unique to each key (i.e., two sets) and those that were common to both keys. Total keyed scores for each set were then validated. We also ran a second kind of cross-validation to find how well each key works in predicting the the length of service on which it was not developed. That is, we calculated validities for the long-tenure key on the short criterion period and for the short-tenure key on the long criterion period.

Results

Table 1 shows how many items were keyable at both tenures and how many were uniquely keyable at only one. Validities for these sets of items and for the total set that was keyable for each condition (unique plus common) are

Table 1
Validities for males of sets of items that were keyable at only the short tenure, only the long tenure, and at both

Criterion	Tenure at Which Items Were Keyed: Items (n) — Short: Total (145)	Short: Unique (23)	Short: Common (122)	Long: Total (181)	Long: Unique (59)	Long: Common (122)
Short						
Development sample	.25	.17	.24	.18	.10	.21
Holdout sample	.19	.14	.19	.18	.11	.19
Long						
Development sample	.22	.09	.23	.31	.27	.30
Holdout sample	.18	.11	.18	.26	.25	.24

Note. The critical value for a difference between two independent correlation coefficients, one for the development sample (n = 4,594) and one for the holdout sample (n = 3,059), is .046 ($p < .05$, two-tailed).

Table 2
Validities by sample and by tenures for keying and for validating; rates of success

Group	N	Tenure on Which the Items Were Keyed: Short (145 Items) — Criterion length: Short	Criterion length: Long	% Stay	Long (181 Items) — Criterion length: Short	Criterion length: Long	% Stay
All Males							
Development	4,594	.25	.22	.87	.18	.31	.75
Holdout	3,059	.19	.18	.86	.18	.26	.75
Females							
Sample 1	1,077	.14	.11	.80	.14	.15	.77
Sample 2	686	.19	.15	.79	.16	.16	.76
Non-graduate males							
Sample 1	743	.20	.12	.79	.13	.19	.56
Sample 2	507	.20	.11	.80	.12	.19	.58
Graduate males							
Sample 1	3,888	.22	.20	.88	.17	.27	.79
Sample 2	2,515	.21	.20	.88	.16	.25	.79

also given. Validities and cross-validities at both the tenure for keying and the other tenure are given in Tables 1 and 2. Table 2 gives validities and success rates for various groups of cases: all males, females, graduate males, and non-graduate males.

Table 3
Descriptive statistics on development and holdout samples as a function of the tenure for keying items and the criterion for validating total scores

	Tenure on Which the Items Were Keyed							
	Short				Long			
Criterion	N	m	sd	t[a]	N	m	sd	t[a]
Short								
Development sample								
Stayers	3,993	252.5	15.6	14.01	3,993	257.3	20.3	11.27
Leavers	601	240.2	20.6		601	245.9	23.6	
Holdout sample								
Stayers	2,629	252.3	15.5	9.43	2,629	256.9	19.9	9.77
Leavers	430	243.1	19.2		430	246.7	21.3	
Long								
Development sample								
Stayers	3,457	253.0	15.5	14.02	3,457	259.6	19.4	20.64
Leavers	1,137	244.3	18.9		1,137	244.5	21.9	
Holdout sample								
Stayers	2,306	252.7	15.4	9.55	2,306	258.5	19.5	14.88
Leavers	753	245.7	18.1		753	246.2	20.1	

[a] $p = .0001$

Table 3 gives mean total scores and standard deviations for stayers and leavers in the development and cross-validation samples and results of t-tests on their means. These results are given for the cases where items were keyed and validated on the same and on different time periods.

Discussion

In five different respects, these data support the hypothesis that tenures longer than the traditional six months are better for keying and validating biodata. First, a full 46% of attrition in this sample occurred after the six-month point. Thus, a key developed at that point is degraded by the presence of almost half of the leavers in the successful criterion group. Second, while over half of the valid items are keyable at both the short and long tenures, more than twice as many are uniquely keyable at the longer one (59 vs 23). Thus a longer instrument results from extending the period for keying.

Third, validity and cross-validity are higher when items are keyed and validated on the longer period. It is true that congruence in the tenures for keying and validating (i.e., either Short key with Short criterion or Long key with Long Criterion) produce the highest sets of validities here; but still the original validity in the Short-Short condition (.25) does not exceed the cross-validity in the Long-Long condition (.26). Similarly, the Long key for the common items has as high a cross-validity for the Short criterion as does the Short key for any set of items, while it has a higher validity at the Long criterion than any set of items with the Short key does.

Fourth, shrinkage of cross-validities is less for item sets that are keyed at the long tenure. In Table 1 the median shrinkage for Short keys is .045 while for Long keys it is .02. Finally, the largest mean differences in total score, both in terms of keyed points and in t-value are for keying and validating at the longer tenure (Table 3).

The data in Table 1 support one other optimistic conclusion. Although the sets of unique items have fairly low validities for the criterion on which they were not keyable, the 59 items which were significant at only the long tenure have a good validity and cross-validity for the longer criterion. Among the highest validities in that table are those that come from this set of about one-third of the items that are useful over that longer period. This finding implies that there may be enough valid items to produce several test forms of satisfactory validity. Among other things, the issue of how to assign items to forms needs to be addressed.

A second topic for further research is that of possible differences in early and late leavers. If found, any such differences might help to explain differences between leavers and stayers. A comparison of the content of the two unique sets of items may yield some hypotheses on this issue.

Although these results confirm the statistical superiority of keying and validating on longer tenures, that practice has a cost: that of delaying implementation of the instrument while the criterion matures. One question for further research is how to balance the benefits of high validity with those of early implementability so as to maximize the net benefit.

The results for females and for non-graduate males are not as positive as for men overall. Whether a good unisex scoring key could be developed remains to be seen. From the the percents of stayers in Table 2, attrition seems to be a somewhat different process in males and females: unlike males' attrition, almost all of females' occurs in the first six months.

Even though the samples of females and non-graduate males are large in absolute numbers, they may not be large enough in these data to produce stable performance in a biodata instrument. Two aspects of the military research setting make results from validations of non-cognitive predictors relatively unstable. First, attrition is managed, and policy on acceptable levels thereof varies over the years. Thus the criterion is driven by at least one force that is not tightly connected with the characteristics of the examinees. Second, the characteristics of the applicant and accession pools also change over the years. For example, a decade ago about half of accessions were non-graduate males; now the rate is around 10%. These facts make

it important to use large, stable samples for developing keys.

Previous attempts to evaluate the validity of MAP in the operational setting have found validities to be much lower than in the research setting (Walker, 1984). Unlike the present research, past work in developing scoring keys has not cross-validated them. The robust cross-validities for the long-long condition here give reason to believe that the keys developed here would retain a good level of validity if put into operation. Even with that assumption, further research on the rates of accurate and inaccurate selection decisions to be expected should be carried out to see whether the instrument is likely to be cost-effective.

References

Cascio, W. F. (1982). Applied psychology in personnel management. Reston, VA: Reston.

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

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

Riegelhaupt, B. J. & Bonczar, T. P. (1985, October). The utility of educational and biographical information for predicting military attrition. Proceedings of the 27th Annual Meeting of the Military Testing Association. San Diego, CA

Walker, C. B. (1984, November). Validating the Army's Military Applicant Profile against an expanded criterion space. Proceedings of the 26th Annual Meeting of the Military Testing Association. Munich, FRG.

Walker, C. B. (1985, October). Three variables that may influence the validity of biodata. Proceedings of the 27th Annual Meeting of the Military Testing Association. San Diego, CA.

ASVAB VALIDITIES USING IMPROVED JOB PERFORMANCE MEASURES

Lauress L. Wise
Jeffrey J. McHenry
American Institutes for Research

Paul G. Rossmeissl
U.S. Army Research Institute

Scott H. Oppler
American Institutes for Research

Presented on Symposium,

"Project A Concurrent Validation: Preliminary Results"

At the Annual Conference of the
Military Testing Association
Mystic, Connecticut

November 1986

ASVAB VALIDITIES USING IMPROVED JOB PERFORMANCE MEASURES

Lauress L. Wise, Jeffrey J. McHenry - American Institutes for Research
*Paul G. Rossmeissl - U.S. Army Research Institute
**Scott H. Oppler - American Institutes for Research

Project A job performance measures are unique in their combination of depth (work samples, ratings, knowledge tests, and administrative measures) and breadth (19 very diverse jobs). This paper examines the validity of the Army's ASVAB Aptitude Area (AA) Composites for predicting job performance as assessed by these new measures. Project A performance measures have been organized into five constructs (Wise, Campbell, McHenry, Hanser, 1986). Four of these constructs (General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing) are the same for each Military Occupational Specialty (MOS). Armed Forces Qualifying Test (AFQT) scores and other selection criteria (e.g. high school graduation, moral and physical requirements) are designed to predict performance on these common constructs. The fifth construct, Core Technical Proficiency (CTP), covers aspects of job performance unique to each MOS. AA scores, used as job specific selection criteria, are appropriately validated against this construct.

In addition to evaluating current AA composites, we identified specific alternative composites. We did not identify alternative composites for every MOS, since we had data for only 19 of the more than 250 entry-level MOS. Instead, we identified alternative composites for each cluster of jobs that currently use the same AA composite. In this paper, we only considered redefining the existing composites. We did not consider changing the assignment of MOS to specific composites.

Methods

Current forms of the ASVAB generate nine subtest scores: General Science (GS), Arithmetic Reasoning (AR), Verbal (VE combining Work Knowledge and Paragraph Comprehension), Coding Speed (CS), Numerical Operations (NO), Auto/Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). AA composites are defined as unweighted sums of four or fewer of the standardized subtest scores. There are 255 such possible composites (126 using four subtests, 84 using three, 36 using two, and 9 using a single subtest). We evaluated all of them.

Project A Concurrent Validation (CV) data were used in evaluating the current composites. The CV data included the new job performance measures applied to over 9,000 soldiers in 19 different MOS. Table 1 shows CV sample sizes by MOS and race and gender and also the ASVAB subtest and the CTP criterion means and standard deviations.

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract Number MDA903-82-C-0531. Statements expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

* Dr. Rossmeissl is now with Hay Systems, Inc. in Washington, D.C.
** Mr. Oppler has returned to graduate work at the University of Minnesota.

Four separate criteria were used in evaluating current and alternative composites: (1) predictive validity, (2) fairness to Blacks and females, (3) classification efficiency, and (4) face validity. Each is described briefly before proceeding to a discussion of the results.

Predictive Validity. The correlation of each composite with the CTP score was adjusted for restriction of range due to explicit selection. A multivariate correction due to Lawley (Lord & Novick, 1968, p. 146) was used with each of the ASVAB subtests treated as a separate selection variable. The result was used as the measure of predictive validity. No adjustment was made for "shrinkage" in cross-validation since separate regression coefficients were not estimated. For evaluation of the current composites, this is entirely appropriate. Because we did pick among a large number of alternative composites on the basis of the data at hand, some shrinkage should be expected for the alternatives that appear most extreme. Conventional shrinkage formulas do not handle this situation, so our best approach is to be somewhat conservative in adopting new alternatives to the existing composites.

Fairness to Blacks and Females. Separate regression equations were computed by race and gender where there were at least 50 examinees. Both slope and intercept differences were identified. A single overall measure of the difference in the separate equations was defined in terms of the expected criterion difference for an AA score of 100 (the estimated 1980 norm population mean.) Since selection cutoffs varied between 85 and 110 for the MOS in question, a score of 100 was selected as being in the heart of the critical region for evaluating the selection fairness of alternative composites. Differences in the prediction equations at points significantly below or above this value would have little impact on determination of applicant qualification. The difference in predicted values was converted to a t score by dividing by the standard error of the estimate of the difference (Pothoff, 1964).

Classification Efficiency. The Brogden index, defined as the square root of the average validity times the square root of one minus the average of the intercorrelations among the composites was used as a measure of classification efficiency. This statistic is an indicator of the accuracy of predictions of differences in an individual's expected performance across jobs.

Face Validity. The final evaluation factor was face validity. Face validity is not easily quantifiable, but is more appropriately used as a check of the "reasonableness" of the results. It is our attempt to check purely empirical results against some conception of theory. We would be uncomfortable, for example, with results indicating that AS is an important predictor for clerical jobs, but quite comfortable with AS as an important predictor for vehicle mechanics.

Results

Table 2 shows validities, Brogden indices (Clss. Eff.), and, where appropriate, race and gender t statistics for each contending AA composites. Separate statistics are shown for each applicable MOS and unweighted averages of the validities and t statistics are shown for the cluster as a whole. Each row of statistics corresponds to a different composites. The first row gives statistics for the current composite. Rows with data on alternative composites are labelled A1 through A9. Data are also shown for the CL and SC composites replaced in 1984 after our prior analyses (McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984) with

the previous composites labelled PR. Where some other of the current composites has a higher average validity than the operational composite the cluster, data are shown in rows that are labelled according to the other composite. The results presented in Table 2 are discussed separately for each of the current AA composites.

Clerical (CL). The current CL composite has a higher average validity than any alternative. It does, however, underpredict female performance in the two clerical specialties where separate predictions were generated. The addition of either NO or CS significantly reduces the underprediction for females without significantly reducing validity. Adding NO reduces underprediction the most, while adding CS has the greatest face validity and results in slightly greater classification efficiency. A slightly different pattern was found for 76W. The addition of AS increases validity for predicting 76W performance, while decreasing validity for predicting 71L and 76Y performances. Notwithstanding these differences, the current and primary alternative CL composites predict performance in all three clerical MOS quite well.

Combat (CO). The current CO has high validity each of the MOS examined. Some gain in validity would be realized by substituting GS for CS and, perhaps, also swapping MK for AR. The inclusion of GS would improve prediction in all three MOS. The greater contribution of GS also is rational in light of increasing technical sophistication in the systems used in combat specialties. Adding GS would also reduce the small degree of overprediction of the performance of Blacks.

Electronic (EL). The current EL composite does quite well for the one EL specialty examined. Substitution of NO for one or both of the quantitative subtests would increase both predictive validity and classification efficiency, but not to any practical extent.

Field Artillery (FA). Neither the current FA nor any alternative appears to have a very high validity for predicting 13B performance. Consideration of alternative composites is motivated by the fact that several other current composites have higher validities for predicting 13B performance than the current FA composite. Substitution of NO and AS for CS and MK would yield the most significant gains. Such substitution also significantly reduces overprediction for Blacks.

General Maintenance (GM). Very high validities were found for the current GM composite for both 51B and 55B. Very slight gains might result from substituting VE for EI or from simply dropping EI, but these gains would be offset by small increases in overprediction of Blacks' performance and slightly lower classification efficiency estimates.

Mechanical Maintenance (MM). High validities were found for the current MM composite in predicting both 63B and 67N performance. Small gains in the prediction of 63B performance and increased classification efficiency would result from dropping the NO subtest.

Operators/Food (OF). The OF results closely parallel the CL results. Female performance is significantly underpredicted for 94B. Another specialty, 64C, shows a somewhat different pattern of validities, with AS again (and not surprisingly) adding significantly to the predictive validity of this one specialty. In fact, the same composites appear optimal for both the CL and OF MOS -- AR+VE+MK+NO for 16S and 94B (as for 71L and 76Y) and AR+VE+MK+AS for 64C (as for 76W). Substituting AR and MK for AS and MC would significantly reduce underprediction of

female performance for 94B while increasing overall validity.

Surveillance and Communication (SC). A high predictive validity was found for the current SC composite. Some gain in validity, along with a slight increase in classification efficiency, would result if MC were replaced by NO. This would lead to a small increase in the underprediction of performance for Blacks. If MK were also substituted for AR, the same gains in validity and classification efficiency could be obtained along with a decrease in underprediction of Blacks' performance.

Skilled Technical (ST). The current ST is a true Army composite -- it is all that it can be. It has a higher average validity than any possible alternative, and it shows no significant differences in the prediction of performance for Blacks and females.

Summary

The Army's existing AA composites were found to have very high validity for predicting job-specific performance as assessed with the Project A measures. A few changes to the existing AA composites to improve validity or reduce gender differences were identified for further consideration. Specific recommendations are:

CL: Add NO to reduce gender differences.
CO: Replace GS with CS to increase validity/reduce race differences.
FA: Replace CS and MK with NO and AS to increase validity.
MM: Drop NO to increase validity.
OF: Replace NO and MC with AR and MK to increase validity.
Reassign 94B (and similar MOS) to CL to reduce gender differences.
SC: Replace AR and MC with MK and NO to increase validitiy and reduce race differences.

Recommendations for further analyses include: (1) investigation of criterion factors associated with low ASVAB correlations for the 13B measures and significant gender differences for 71L and 94B and (2) evaluation of alternative assignment of MOS to composites, particularly for the CL and OF composites.

References

Lord, F. & Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

McLaughlin, D., Rossmeissl, P., Wise, L., Brandt, D., & Wang, M. (1984). Validation of current and alternative ASAB Area Composites, based on training and SOT information on FY 1981 and FY 1982 Enlisted Accessions (Technical Report 651). Alexandria, VA: U. S. Army Research Institute.

Potthoff, R. (1964). On the Johnson-Newman technique and some extensions thereof. Psychometrika, 29, 241-245.

Wise, L., Campbell, J., McHenry, J., & Hanser, L. (1986, August). A latent structure model of job performance factors. Paper presented at the annual meeting of the American Psychological Association.

Table 1. Descriptive Statistics

MOS	COMP	GROUP	N	MEAN										STANDARD DEVIATION									
				CTP	GS	AR	VE	CS	MO	AS	MK	MC	EI	CTP	GS	AR	VE	CS	MO	AS	MK	MC	EI
11B: INFANTRY	CO	ALL	491	514	529	539	519	525	515	557	515	551	533	80	80	73	63	64	65	76	78	75	73
12B: COMBAT ENG	CO	ALL	544	509	506	527	496	510	499	555	502	539	522	96	86	70	71	66	60	81	77	83	76
		BLACK	108	453	433	482	440	495	489	479	460	478	473	78	66	47	58	65	54	62	47	58	57
		WHITE	385	529	533	542	519	514	501	584	515	559	539	94	77	70	62	66	61	65	80	81	73
13B: CANNON CREW	FA	ALL	464	510	487	519	488	516	497	514	495	509	502	85	87	69	70	61	65	91	66	83	78
		BLACK	168	485	438	491	456	516	493	458	478	466	467	84	76	60	68	59	68	71	55	70	67
		WHITE	250	528	528	544	518	518	501	563	507	546	533	82	73	65	57	61	63	74	71	74	69
16S: MANPAD CREW	OF	ALL	338	516	509	519	505	527	498	548	495	531	527	94	81	79	66	64	76	81	77	84	76
		BLACK	89	494	449	469	460	540	489	481	464	477	484	78	77	57	52	60	75	67	55	67	60
		WHITE	232	524	534	541	524	522	500	578	510	553	546	99	71	77	62	65	76	69	81	79	74
19E: ARMOR CREW	CO	ALL	394	514	527	536	513	515	506	567	515	549	535	75	84	73	69	66	67	79	77	78	80
		BLACK	71	469	459	497	465	499	483	497	477	488	474	69	77	56	64	65	67	66	63	63	63
		WHITE	297	524	548	547	530	517	511	588	525	568	553	75	75	74	60	65	66	70	76	73	76
27E: TOW/DRG REP	EL	ALL	123	505	540	552	524	518	504	561	532	548	560	101	66	62	58	69	68	75	69	72	70
31C: RADIO/TTY	SC	ALL	289	508	518	540	521	554	547	547	521	527	514	85	76	72	59	54	60	80	79	86	79
		BLACK	74	488	461	494	494	564	557	498	493	479	482	69	68	71	60	44	66	70	63	77	64
		WHITE	204	513	538	555	532	550	542	565	529	543	525	89	68	66	56	56	56	77	82	84	82
51B: CRPNT/MSNRY	GN	ALL	69	513	508	510	497	505	481	555	491	536	533	101	72	72	60	70	66	70	67	76	64
54E: NPC SPEC	ST	ALL	340	507	540	543	529	517	503	543	533	543	531	99	71	73	57	70	69	82	74	72	76
		BLACK	84	466	505	516	515	508	482	485	516	500	493	98	66	69	55	70	72	63	64	55	68
		WHITE	223	522	558	554	541	520	511	571	538	562	549	95	64	74	52	71	67	74	76	70	71
55B: AMMO SPEC	GN	ALL	203	507	497	495	475	491	476	526	481	490	523	97	64	65	62	64	68	69	60	76	57
		BLACK	75	472	477	469	458	492	475	486	470	451	516	99	48	56	53	61	69	52	43	56	44
		WHITE	112	531	513	513	486	491	475	556	486	519	527	89	69	63	65	68	70	66	70	78	64
63B: VEHCLE MECH	MN	ALL	478	513	506	528	496	520	509	579	501	543	536	76	78	71	62	63	59	78	69	79	65
		BLACK	78	464	445	478	456	520	491	510	476	479	503	72	64	59	64	61	63	70	54	57	52
		WHITE	374	526	522	541	507	519	513	598	508	559	546	70	72	69	57	63	59	69	71	75	66
64C: MOTOR TRANS	OF	ALL	507	510	486	498	481	513	499	548	483	522	509	72	75	76	63	65	67	75	68	76	71
		BLACK	121	487	444	456	450	523	492	498	456	471	471	73	65	60	54	61	69	70	54	65	74
		WHITE	358	520	502	513	493	508	501	568	493	541	523	66	71	77	62	66	65	68	69	72	66
		FEMALE	52	495	485	503	520	554	559	464	490	480	454	71	73	78	55	65	67	65	61	72	54
		MALE	455	512	486	498	477	509	492	558	483	526	515	72	75	76	63	63	64	70	69	75	70
67N: HELCPTR REP	MN	ALL	238	510	567	567	546	550	531	613	550	601	582	93	60	59	47	53	63	54	67	54	57
71L: ADMIN CLERK	CL	ALL	427	506	493	528	514	562	552	476	515	484	481	87	82	72	59	49	61	79	75	79	69
		BLACK	159	491	464	499	495	563	535	444	498	454	464	81	74	65	59	45	63	61	69	71	55
		WHITE	235	516	518	548	531	560	560	502	528	505	496	89	79	70	51	52	58	84	75	79	74
		FEMALE	237	524	486	519	522	566	561	447	508	461	465	72	73	67	49	50	63	64	66	68	52
		MALE	190	483	502	539	505	558	540	514	524	512	501	98	91	76	68	48	57	82	84	83	82
76W: PETRO SPPLY	CL	ALL	339	519	479	511	494	536	512	508	491	500	498	95	90	74	69	54	65	99	72	91	81
		BLACK	139	476	430	472	463	539	500	447	461	444	461	88	73	63	65	52	64	73	60	66	67
		WHITE	174	551	521	539	522	535	518	560	514	548	530	88	85	69	60	55	64	90	73	84	78
76Y: UNIT SUPPLY	CL	ALL	444	516	489	518	500	550	531	496	507	496	496	93	85	74	67	51	58	86	75	84	78
		BLACK	169	487	442	479	473	553	518	455	473	453	463	90	69	62	60	46	54	71	60	65	63
		WHITE	231	536	528	547	524	547	538	532	530	530	524	93	76	71	60	56	61	83	78	83	78
		FEMALE	75	519	463	501	492	569	551	429	494	448	453	84	73	62	59	48	61	57	71	72	62
		MALE	369	516	494	522	501	546	527	510	509	506	504	95	87	76	68	51	57	84	76	83	78
91A: MEDIC SPEC	ST	ALL	392	514	547	544	540	525	520	528	530	543	524	79	62	64	46	69	70	82	71	70	68
		BLACK	91	486	519	512	521	519	508	486	511	495	496	72	50	58	42	74	62	70	65	54	58
		WHITE	260	525	562	555	550	527	524	548	538	560	534	80	61	64	42	68	71	80	72	68	70
		FEMALE	116	513	532	545	542	550	549	465	543	504	475	81	59	59	48	58	65	66	66	64	52
		MALE	276	515	554	544	539	514	508	555	525	559	544	79	63	66	46	71	68	73	72	67	63
94B: FOOD SERVCE	OF	ALL	368	526	496	515	503	533	510	516	495	510	503	90	80	77	63	63	69	82	72	76	75
		BLACK	124	493	449	466	471	534	501	469	463	464	471	77	70	58	56	60	72	63	51	56	66
		WHITE	222	546	524	543	524	532	517	546	515	536	524	94	74	73	58	66	66	79	73	75	74
		FEMALE	78	553	474	499	513	562	546	448	489	467	446	79	80	65	53	57	82	64	64	64	59
		MALE	290	519	502	519	501	526	501	534	497	522	518	92	79	80	65	63	62	77	74	75	72
95B: MIL POLICE	ST	ALL	597	504	562	554	542	530	519	573	537	571	550	74	53	60	42	62	62	68	61	58	62

Table 2. Validity, Cultural Fairness, and Classification Efficiency Indicators for Current and Other ASVAB Composites

Current/Other Composites	Avg. Val	Avg Δ Race	Avg Δ Sex	Class Eff*	Val	Δ by Race	Δ by Sex	Val	Δ by Race	Δ by Sex	Val	Δ by Race	Δ by Sex
CL: CLERICAL					**71L: ADMIN SPEC**			**76W: PETRO SPPLY**			**76Y: UNIT SUPPLY**		
CL: AR+VE+MK	.661	-2.2	16.1	.231	.64	.6	20.4	.67	-5.8		.67	-1.4	11.8
PR: VE+NO+CS	.578	-5.7	3.1	.248	.59	-.2	5.6	.55	-12.8		.60	-4.3	.5
A1: AR+VE+NO+MK	.656	-3.1	6.7	.232	.65	.4	10.6	.65	-7.8		.67	-2.0	2.9
A2: AR+VE+CS+MK	.656	-2.2	8.1	.233	.65	1.6	11.4	.65	-7.0		.67	-1.1	4.9
A3: AR+VE+AS+MK	.655	-.5	22.2	.222	.60	1.0	32.2	.70	-2.0		.67	-.4	12.3
CO: COMBAT					**11B: INFANTRYMAN**			**12B: COMBAT ENG**			**19E: ARMOR CREW**		
CO: AR+CS+AS+MC	.617	-3.2	-	.231	.66	-	-	.64	-3.5	-	.55	-3.0	-
A1: GS+AS+MK+MC	.648	-1.9	-	.229	.67	-	-	.67	-2.9	-	.60	-1.0	-
GM: GS+AS+MK+EI	.641	-2.5	-	.230	.67	-	-	.67	-3.5	-	.58	-1.5	-
A2: GS+MK+AS	.643	-2.4	-	.230	.67	-	-	.67	-3.3	-	.59	-1.4	-
EL: ELECTRONIC					**27E: TOW/DRGN REP**								
EL: GS+AR+MK+EI	.779	-	-	.231	.78	-	-						
A1: GS+NO+EI	.791	-	-	.235	.79	-	-						
A2: GS+NO+MK+EI	.791	-	-	.232	.79	-	-						
FA: FIELD ARTILLERY					**13B: CANNON CREW**								
FA: AR+CS+MK+MC	.341	-8.4	-	.231	.34	-	-						
A1: GS+NO+AS+MC	.383	-3.1	-	.227	.38	-	-						
A2: AR+NO+AS+MC	.381	-3.8	-	.227	.38	-	-						
GM: GENERAL MAINTENANCE					**51B: CRPNT/MSNRY**			**55B: AMMO SPEC**					
GM: GS+AS+MK+EI	.785	-5.0	-	.231	.81	-	-	.76	-5.0	-			
A1: GS+VE+AS+MK	.798	-6.3	-	.229	.84	-	-	.75	-6.3	-			
A2: GS+AS+MK	.791	-6.4	-	.230	.84	-	-	.74	-6.4	-			
A3: GS+AR+VE+AS	.789	-4.5	-	.228	.82	-	-	.76	-4.5	-			
A4: GS+CS+AS+MK	.789	-10.0	-	.229	.86	-	-	.72	-10.0	-			
MM: MECHANICAL MAINTENANCE					**63B: VEHCLE MECH**			**67N: HELCPTR REP**					
MM: NO+AS+MC+EI	.729	-4.7	-	.231	.66	-4.7	-	.80	-	-			
A1: AS+MC+EI	.745	-4.5	-	.240	.69	-4.5	-	.80	-	-			
A2: GS+AS+MC+EI	.742	-4.4	-	.233	.68	-4.4	-	.81	-	-			
A3: AS+MK+MC+EI	.739	-5.6	-	.229	.67	-5.6	-	.81	-	-			
A4: AR+MC+AS+EI	.739	-4.3	-	.230	.67	-4.3	-	.81	-	-			
A5: GS+AS+MC	.738	-3.9	-	.234	.67	-3.9	-	.81	-	-			
A6: AS+MC	.733	-3.5	-	.244	.68	-3.5	-	.79	-	-			
OF: OPERATORS/FOOD					**16S: MANPAD CREW**			**64C: MOTOR TRANS**			**94B: FOOD SERVCE**		
OF: VE+NO+AS+MC	.538	-1.0	8.4	.231	.44	.9	-	.52	-1.4	-4.6	.65	-2.5	21.3
A1: AR+VE+AS+MK	.571	.8	9.0	.228	.51	3.0	-	.53	-.5	-14.1	.68	-.2	32.1
A2: GS+AR+AS+MK	.568	.5	10.7	.228	.50	2.9	-	.54	-.1	-4.8	.67	-1.5	26.2
A3: AR+AS+MK	.567	-.2	12.3	.230	.49	2.1	-	.54	-1.1	-2.3	.66	-1.4	26.9
A4: GS+AR+MK	.561	-1.1	10.0	.232	.52	2.2	-	.49	-3.6	-15.5	.68	-1.9	35.5
A5: GS+AR+VE+MK	.561	-.8	13.3	.231	.52	2.7	-	.48	-3.7	-17.6	.69	-1.5	44.1
A6: AR+VE+MK	.558	-1.4	13.2	.228	.52	2.0	-	.46	-5.2	-19.0	.69	-1.2	45.4
A7: AR+VE+MK+MC	.566	-.4	6.4	.234	.50	1.7	-	.51	-2.4	-24.7	.69	-.6	37.6
A8: AR+VE+NO+MK	.548	-4.8	-1.8	.236	.51	-.1	-	.44	-10.8	-16.5	.70	-3.4	13.0
A9: AR+VE+CS+MK	.546	-3.2	2.3	.236	.51	.2	-	.44	-6.9	-14.7	.70	-2.9	19.3
EL: GS+AR+MK+EI	.558	-.8	9.4	.228	.50	2.1	-	.51	-2.0	-7.1	.66	-2.6	25.8
ST: GS+VE+MK+MC	.557	.6	7.1	.228	.50	-1.4	-	.51	1.9	-16.9	.66	-3.0	31.1
FA: AR+CS+MK+EI	.555	-2.9	6.3	.230	.49	-.7	-	.49	-5.1	-22.6	.69	-3.0	35.3
SC: SURVEILLANCE & COMMUNICATION					**31C: RADIO/TTY OP**								
SC: AR+VE+AS+MC	.693	1.9	-	.231	.69	1.9	-						
PR: VE+NO+CS+AS	.701	.5	-	.232	.70	.5	-						
A1: AR+VE+NO+AS	.729	2.4	-	.233	.73	2.4	-						
A2: VE+NO+AS+MK	.729	.9	-	.233	.73	.9	-						
A3: AR+VE+NO+EI	.728	1.2	-	.234	.73	1.2	-						
A4: GS+AR+NO+EI	.727	2.0	-	.232	.73	2.0	-						
ST: SKILLED TECHNICAL					**54E: NBC SPEC**			**91A: MEDIC SPEC**			**95B: MIL POLICE**		
ST: GS+VE+MK+MC	.683	-1.5	.1	.231	.69	-1.6	-	.73	-1.3	.1	.63	-	-
A1: GS+CS+AS+MK	.679	-1.1	.5	.231	.67	-1.5	-	.75	-1.5	.5	.62	-	-

VALIDATION ANALYSIS FOR NEW PREDICTORS

John P. Campbell

Human Resources Research Organization

Presented at a Data Analysis Workshop of the
Committee on Performance of Military Personnel

Baltimore

December 1986

This is a working paper prepared for a conference on performance measurement sponsored by the Committee on the Performance of Military Personnel of the National Research Council. It reports some of the initial selection test validation data collected by the Army's Selection and Classification Project (Project A). These data should still be viewed as preliminary and should not be cited, quoted, or distributed.

After briefly summarizing the objectives of the project, the basic data collection design, and the steps used in the development of the new selection tests and criterion measures, the intitial validity results from the concurrent sample will be presented.

Much of the background information that outlines the objectives, sample characteristics, selection test development, and performance measure development is taken from a 1986 APA paper (Campbell, 1986) that gave an overview of the entire project. Most of the results that are presented were generated by analyses done since that time.

Overall Project A Objectives

Project A is directed at multiple operational and research objectives. The major ones are shown in Table 1.

The current Army selection/classification system for enlisted personnel screens 300 to 400 thousand people each year, selects 120-140 thousand of them, and assigns each individual to one of approximately 275 entry level positions. The primary selection instrument is the Armed Services Vocational Aptitude Battery (ASVAB) which currently has ten subtests and composites of subtests developed for different categories of occupational specialties. Cutting scores have been established for each job, or Military Occupational Speciality (MOS), and if the individual is above the cutting score on the appropriate ASVAB composite, assignments are made on the basis of Army needs, training space availability, and individual preferences. A system of bonuses is currently in use to influence individual preferences in the direction of Army needs.

The mandate of Project A is to develop an experimental battery of new selection/classification instruments, validate them against appropriate measures of job performance, assess their collective differential validity for making classification decisions, and provide the information necessary for conducting "what if" games with differential weights for job assignments, changes in cutting scores, quotas, etc. The latter activity would be carried out in conjunction with the assignment algorithms developed by Project B.

In the course of trying to meet this mandate, Project A has taken a broad approach. For example, we have tried to provide a systematic description, in a taxonomic sense, of the universe of information that is potentially useful for making predictions of future job performance and to develop a model of its latent structure. Similarly the Project has tried to develop a general latent structure model of job performance for entry level skilled jobs, at least as they are represented by the population of jobs performed by enlisted personnel in the U. S. Army.

Table 1

Army Selection and Classification Project

Operational Objectives

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

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

3) Develop and validate new selection and classification measures.

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

5) Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility.

Research Objectives

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

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

3) Investigate the construct validity of the "method" variance in job performance measures.

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

5) 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.

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

7) Develop new statistical estimators of classification efficiency.

If the latent structure of job performance and the taxonomic structure of selection/classification prediction information is modeled, and measures are developed to assess the major constructs using samples of soldiers from a representative sample of jobs from a large population of jobs, then a number of interesting questions can be examined systematically. For example, to what extent is differential prediction possible across major components of performance? To what extent is there differential prediction across the major components of the predictor universe? To what extent does validity generalization across jobs depend upon the performance component being assessed? For any differential prediction across race and gender groups, what is the source of such differential regressions in terms of predictor component/performance component combinations? What happens to the overall regression picture when those components are omitted?

Basic Project Design

The basic design of Project A shown in Figure 1 is simply that of a very large test validation study that incorporates several independent data collections.

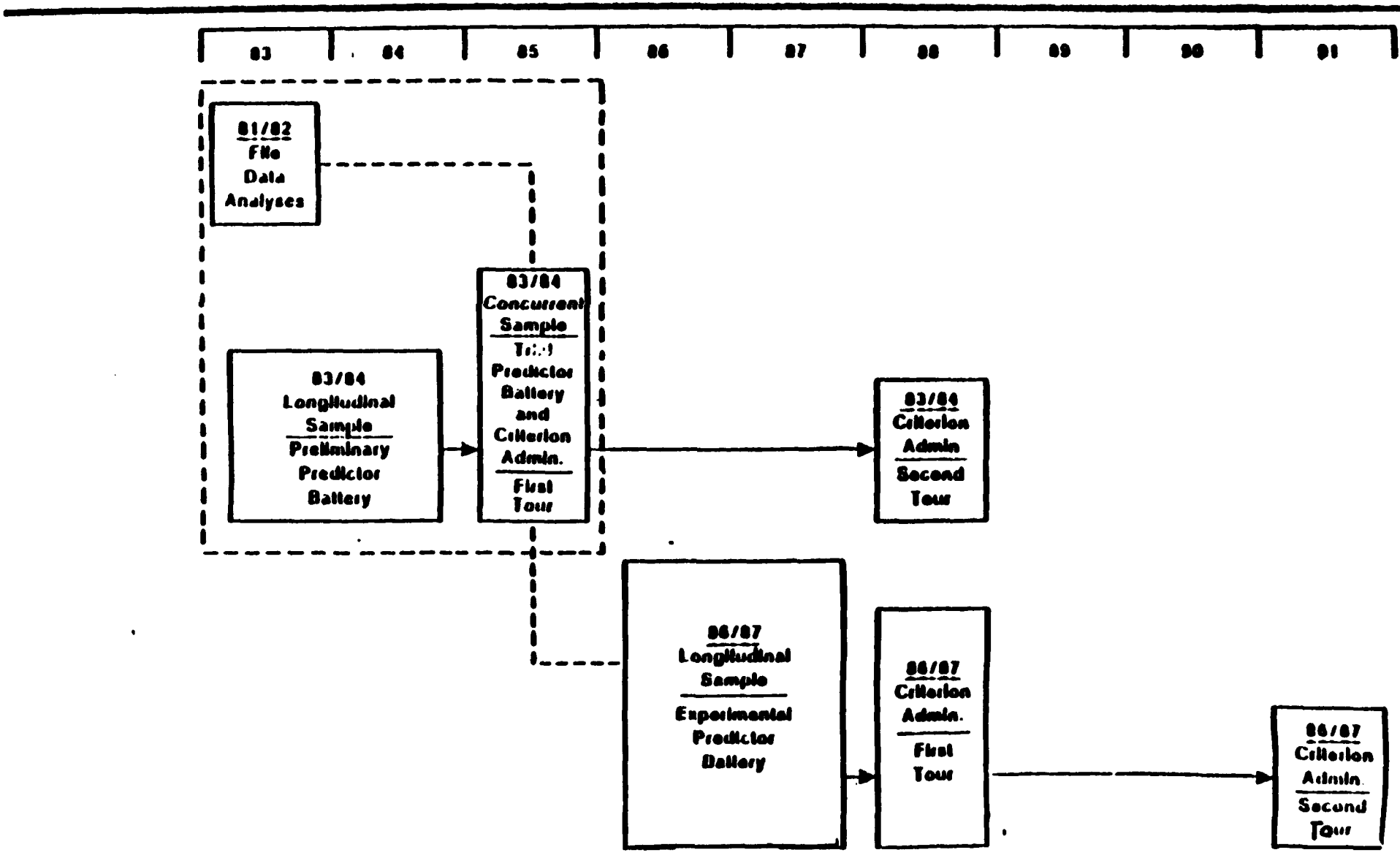
There are four major data files. The first consists of the available computer records for people who joined the Army in 1981 and 1982. The basic data are the ASVAB, the available training school grades, and the Skills Qualification Test (SQT), which is a paper-and-pencil measure of current job knowledge constructed, administered, and scored by the individual's unit command. Complete data were available on at least 100 people for 98 of the 275 MOS.

The three waves of new data collected by the project consist of (1) a longitudinal sample called the preliminary battery sample which is composed of approximately 2,000 recruits in each of 4 MOS, (2) a major concurrent validation sample composed of 400-600 incumbents in each of 19 MOS, and (3) an even larger longitudinal validation sample composed of over 40,000 recruits taken from 21 MOS. Besides providing different kinds of validity information, the three samples were intended to provide the opportunity for multiple revisions of the new predictor battery. The preliminary battery sample was assessed with a four-hour battery of carefully selected off-the-shelf tests to provide a set of marker variables for the project-developed tests. Approximately one-fifth of this sample became a part of the concurrent validation sample, which was the first time the full array of project-developed tests and performance measures were administered together. Each job incumbent in the concurrent validation sample was assessed eight hours each day for two days. The longitudinal validation builds upon the concurrent findings and is designed to yield a sample of 400-600 per MOS after the decay rates for the MOS cohorts have their effect. To produce a sample of 10,000 incumbents at the time of job performance assessment, approximately 45,000 new recruits are being tested on the predictor battery.

The reenlistees from both the concurrent sample (83/84 cohort) and from the longitudinal sample (86/87 cohort) will be followed into their second

FIGURE 1

THE RESEARCH FLOW



tour and assessed with another array of job performance measures. During the second tour the job tasks require a higher level of skill and the supervisor/leadership component becomes much more prominent.

Predictor Development

The standard operating procedure for predictor development in personnel selection research is to do a job analysis first. On the basis of a job analysis, the knowledges, skills, and abilities (KSA) required for successful performance are inferred, and an additional judgment is then made about which KSA are trainable and which must be selected for. We didn't precisely do that in Project A.

Instead, the 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 long series of pilot and field tests. The intent was to develop a predictor battery that was maximally useful for an entire population of jobs and not to tailor-make them for the specific jobs in the sample. The loss in specific prediction accuracy for the jobs in the sample (if any) should be compensated for by the gain in coverage for all other jobs in the population.

The long process of predictor development is represented in Figure 2.

It began with an exhaustive search of the entire personnel selection literature. Research teams were created for cognitive abilities, perceptual and psychomotor abilities, and non-cognitive characteristics such as temperament, 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 reviews and eventually came down to a list of 53 potentially useful predictor constructs. They are listed in Table 2.

A similar, but different, procedure was used to identify a population of performance factors - 72 in all. We then assembled a sample of 35 personnel selection experts and asked them 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 were analyzed for inter-judge agreement, rows and columns were factor analyzed, and the results were compared to analogous information from the empirical literature. Most importantly, however, 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 2.

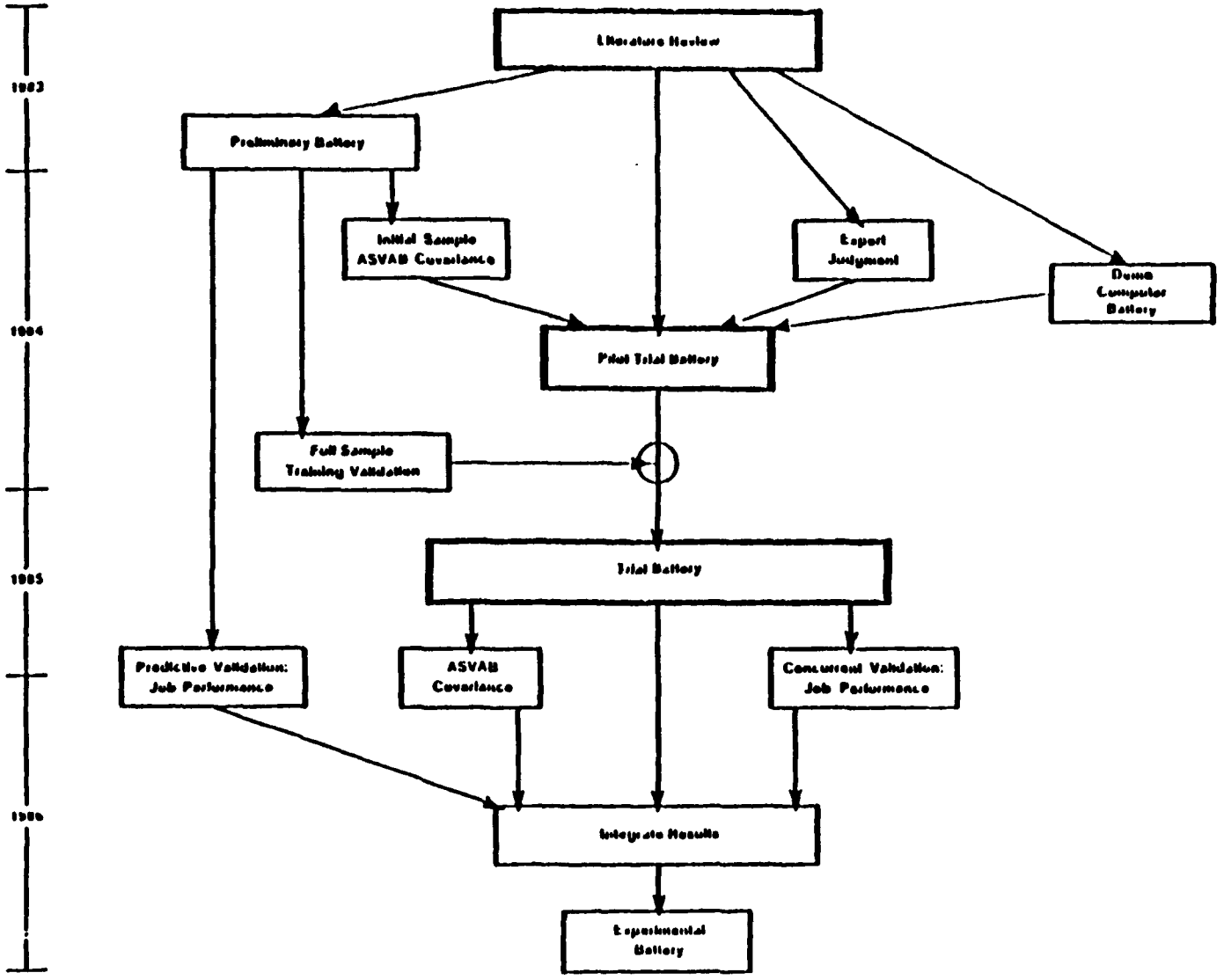


Figure 2. Flow Chart of Predictor Measure Development Activities of Project A.

CONSTRUCTS	CLUSTERS	FACTORS
1. Verbal Comprehension 5. Reading Comprehension 16. Ideational Fluency 18. Analogical Reasoning 21. Omnibus Intelligence/Aptitude 22. Word Fluency	A. Verbal Ability/ General Intelligence	
4. Word Problems 8. Inductive Reasoning: Concept Formation 10. Deductive Logic	B. Reasoning	
2. Numerical Computation 3. Use of Formula/Number Problems	C. Number Ability	COGNITIVE ABILITIES
12. Perceptual Speed and Accuracy	E. Perceptual Speed and Accuracy	
49. Investigative Interests	U. Investigative Interests	
14. Rote Memory 17. Follow Directions	J. Memory	
19. Figural Reasoning 23. Verbal and Figural Closure	F. Closure	
6. Two-dimensional Mental Rotation 7. Three-dimensional Mental Rotation 9. Spatial Visualization 11. Field Dependence (Negative) 15. Place Memory (Visual Memory) 20. Spatial Scanning	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
24. Processing Efficiency 25. Selective Attention 26. Time Sharing	G. Mental Information Processing	INFORMATION PROCESSING
13. Mechanical Comprehension	L. Mechanical Comprehension	MECHANICAL
48. Realistic Interests 51. Artistic Interests (Negative)	N. Realistic vs. Artistic Interests	
28. Control Precision 29. Rate Control 32. Arm-hand Steadiness 34. Aiming	I. Steadiness/Precision	
27. Multilimb Coordination 35. Speed of Arm Movement	D. Coordination	PSYCHOMOTOR
30. Manual Dexterity 31. Finger Dexterity 33. Wrist-Finger Speed	K. Dexterity	
39. Sociability 52. Social Interests	Q. Sociability	SOCIAL SKILLS
50. Enterprising Interests	R. Enterprising Interests	
36. Involvement in Athletics and Physical Conditioning 37. Energy Level	T. Athletic Abilities/Energy	VIGOR
41. Dominance 42. Self-esteem	S. Dominance/Self-esteem	
40. Traditional Values 43. Conscientiousness 46. Non-delinquency 53. Conventional Interests	N. Traditional Values/Conventionality/Non-delinquency	
44. Locus of Control 47. Work Orientation	O. Work Orientation/Locus of Control	MOTIVATION/ STABILITY
38. Cooperativeness 45. Emotional Stability	P. Cooperation/Emotional Stability	

Table 2. Hierarchical Map of Predictor Space

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 lots of 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 74 portable computerized testing stations. After each data collection, revisions were made on the basis of item statistics and expert review. Finally on May 15, 1985, the predictor battery was deemed ready for concurrent validation. That battery, known as the Trial Battery (TB), is listed in Table 3.

Performance Measurement

The goals of training and job performance measurement in Project A 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 multi-trait, multi-method approach), c) develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance, and d) evaluate existing archival and administrative records as possible indicators of job performance.

Given these intentions, the criterion development effort focused on three major methods: hands-on job sample tests, multiple choice knowledge tests, and ratings. The behaviorally anchored rating scale (BARS) procedure was extensively used in the development of the rating methods.

Modeling Performance

The development efforts to be described were guided by a particular "theory" of performance. The basic outline is as follows:

First, job performance really is multi-dimensional. There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization.

Table 3

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

COGNITIVE PAPER-AND-PENCIL TESTS Test Name (Construct Name)	Number of Items
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 Test Name (Construct Name)	Number of Items
Simple Reaction Time (Processing efficiency)	15
Choice Reaction Time (Processing efficiency)	30
Memory Test (Short-term memory)	36
Target Tracking Test #1 (Psychomotor precision)	18
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 #2 (Two hand coordination)	18
Number Memory Test (Number operations)	28
Cannon Shoot Test (Movement judgment)	36

NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES Inventory Name and Constructs	Number of Items
Assessment of Background and Life Experiences (ABLE) Inventory	209
Adjustment	
Dependability	
Achievement	
Physical Condition	
Leadership	
Focus of Control	
Agreeableness/Likeability	
Army Vocational Interest Career Examination (AVOICE)	176
Realistic Interests	
Conventional Interests	
Social Interests	
Enterprising Interests	
Artistic Interests	

Two General Factors

For the population of entry level enlisted positions we postulated that there are two major types of job performance components. The first is composed of components that are specific to a particular job. That is, measures of such components would reflect specific technical competence or specific job behaviors that are not required for other jobs. The second kind of performance factor includes components that are defined and measured in the same way for every job. These are referred to as Army-wide criterion factors.

For the job specific components, we anticipated that there would be a relatively small number of distinguishable factors of technical performance that would be a function of different abilities or skills and which would be reflected by different task content.

The Army-wide concept incorporates the basic notion that total performance is much more than task or technical proficiency. It might include such things as contributions to teamwork, continual self-development, support for the norms and customs of the organization, and perserverance in the face of adversity.

In sum, the working model of total performance with which the project began viewed performance as multi-dimensional within the two broad categories of factors. The job analysis and criterion construction methods were designed to "discover" the content of these factors via an exhaustive description of the total performance domain, several iterations of data collection, and the use of multiple methods for identifying basic performance factors.

Factors vs. a Composite

Saying that performance is multi-dimensional does not preclude using just one index of an individual's contributions to make a specific personnel decision (e.g., select/not select, promote/not promote). As argued by Schmidt and Kaplan (1971) some years ago, it seems quite reasonable for the organization to scale the importance of each major performance factor relative to a particular personnel decision that must be made and to combine the weighted factor scores into a composite that represents the total contribution or utility of an individual's performance, within the context of that decision. That is, the way in which performance information is weighted and combined is a value judgment on the organization's part. The determination of the specific combinational rules (e.g., simple sum, weighted sum, non-linear combination) that best reflect what the organization is trying to accomplish is a matter of research.

A Structural Model

If performance is characterized in the above manner, then a more formal way to model performance is to think in terms of its latent structure, postulate what that might be and then resort to a confirmatory analysis.

Unfortunately, it is true that we simply know a lot more about predictor constructs than we do about job performance constructs. There are volumes of research on the former, and almost none on the latter. For personnel psychologists it is almost second nature to talk about predictors in terms of theories and constructs. However, on the performance side, the textbooks are virtually silent. Only a few people have even raised the issue (e.g., Dunnette, 1963; Wallace, 1965).

Unit vs. Individual Performance

Finally, people do not usually work alone. Individuals are members of work groups or units and it is the unit's performance that frequently is the most central concern of the organization. However, determining the individual's contribution to the unit's performance is not a simple problem. Further, variation in unit performance is most likely a function of a number of factors besides the "true" level of performance of each individual.

For two major reasons, Project A has not incorporated unit effectiveness in its model of performance. First, the project is focused on the development of a new selection/classification system for entry level personnel and is concerned with improving personnel decisions about individuals and not units. The task is to maximize the average payoff per individual selected.

The second major reason is the prohibitive cost. It simply was not possible to develop reliable and valid field exercises for assessing unit performance in a representative sample of jobs within a reasonable time frame. In isolated instances it might be possible to take advantage of regularly scheduled exercises or use existing performance records that a particular unit (e.g., maintenance depot) might keep. However, it proved not possible to obtain such data in any systematic way. Even if it could be done, it would not be easy to establish the correspondence between individual performance and unit effectiveness.

What we have chosen to do is to try to identify the factors, or means, by which individuals contribute to unit performance and to assess individual performance on those factors via rating methods. We also have a certain amount of information on situational and unit characteristics and are attempting to determine how much of the variance in individual performance is accounted for by those characteristics.

Criterion Development

Actual criterion development proceeded from two basic types of information. First, all available task descriptions were used to generate a population of job tasks for each MOS. The principal sources of task description are the Army's periodic job description surveys, which use 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. The two sources describe tasks at a somewhat different level of generality with the occupational survey items being much more specific in nature.

Unfortunately, no textbook or available technology tells us what the specifications of a task description should be for different purposes. We opted for statements which described a complete operation, which had a recognizable beginning and end, and which were relatively independent of other tasks. That is, it is possible to perform Task A without performing Task B. After much editing, revising, and a formal review by a panel of subject matter experts, a population of 130-180 tasks was enumerated for each MOS.

An additional series of expert judgments was then used to scale the relative difficulty and importance of each task and to cluster tasks on the basis of content similarity. Sampling tasks for measurement was accomplished via a kind of Delphi procedure. That is, each member of a team of task selectors was asked to select 30 tasks from the population of tasks such that 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 each revised their choices as they saw fit. Typically, convergence was achieved quickly and the final selection was by consensus. The consensus of the task selection panel was then thoroughly reviewed by the Army command responsible for that particular job.

Standardized job samples, the paper-and pencil job knowledge tests, and numerical ratings scales were then constructed to assess knowledge and proficiency on these tasks. Each measure went through multiple rounds of pilot testing and revision. The job sample tests were fairly elaborate and were composed of multiple stations sometimes spread over an area of football field size. Each task to be tested was broken down into several steps each of which was scored pass/fail.

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

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

Since one major objective was to determine the relationships between training performance and job performance and their differential predictability, if any, a comprehensive training achievement test was constructed for each MOS by carefully matching the content of the program of instruction (POI) with the content of the population of job tasks, and writing items to represent each segment of the match. We were most interested in task content which is taught, and also performed on the job, versus tasks which were performed on the job but not part of the POI. Scores on this latter category of items (when given to trainees) would be a measure of incidental learning.

The correlation of direct learning and incidental learning with job performance, both when initial ability is controlled and when it is not, is of considerable interest.

The final entry in the array of criterion measures was produced by a concerted effort to get what we could from the files or archival records. Potentially at least, there are numerous performance indicators lurking in existing computer records and personnel files. We began by enumerating all possibilities from three major sources of such records.

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

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

Military Personnel Records Jacket (MPRJ) - or more commonly known as the 201 file which is the personnel folder that follows the individual.

We systematically compared these three sources using a sample of 750 people and a standardized information recording form. The 201 file looked the most promising in terms of recency and completeness, but of course it is by far the most expensive to search. (The textbooks never mention these cost-benefit questions.) As a consequence, everyone crossed their fingers and we collected eight archival performance indicators via a self-report questionnaire. That is, people were asked what was in their personnel file as regards letters of commendation, disciplinary actions, etc. Field tests on a sample of 500 people showed considerable agreement between self-report and archival records. Almost all disagreements were in the direction of more frequent self-reports, for both positive and negative things. Further followup questionnaires and interviews suggested that self-report may be the more accurate. Anyway, we used them and their distributions and correlations seemed quite reasonable. The self-report items were combined into four indicators that were actually used as criterion measures.

The complete array of performance measures in the form in which they survived a large scale field study of N = 150/MOS for nine MOS is shown in Table 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 the Batch A (9 MOS) and Batch Z (10 MOS) is that not all criterion measures were developed for each job in Batch Z. Budget constraints dictated that the job-specific measures could only be developed for a limited number of jobs (i.e., Batch A).

Table 4

Summary of Criterion Measures Used in Concurrent Validation Samples[1]

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 (non-commissioned officer) potential.

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

5. A 40-item summated rating scale for the assessment of expected combat performance.

6. Five performance indicators from administrative records. The first four are obtained via self-report and the last one from computerized records.

   - o Total number of awards and letters of commendation.
   - o Physical fitness qualification.
   - o Number of disciplinary infractions.
   - o Rifle marksmanship qualification score.
   - o Promotion rate (in deviation units).

Performance Measures for Batch A Only

7. Job-sample (hands-on) test of MOS-specific task proficiency.

   - o Individual is tested on each of 15 major job tasks.

8. Paper-and-pencil job knowledge tests designed to measure task-specific job knowledge.

   - o Individual is scored on 150-200 multiple choice items representing 30 major job tasks. Fifteen of the tasks were also measured hands-on.

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

10. MOS-specific behaviorally anchored ratings scales. From 7 to 13 BARS were developed for each MOS to represent the major factors that constituted job-specific technical and task proficiency.

Performance Measures for Batch Z Only

11. 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.

Auxiliary Measures Included in Criterion Battery

12. Job History Questionnaire which asks for information about frequency and recency of performance of the MOS-specific tasks.

13. Work Environment Description Questionnaire - a 141-item questionnaire assessing situational/environmental characteristics, leadersnip climate, and reward preferences.

[1]All rating measures were obtained from approximately 2 supervisors and 3 peers for each ratee.

Results From the Concurrent Validation Sample

If all the rating scales are used separately, the MOS-specific measures are aggregated at the task or instructional module level, and the major predictor subscales are used, there are approximately 200 criterion scores and 60-70 predictor scores on each individual.

At this point, a classic argument arises between the empirical keying/ "let's look for all the specific variance we can" types and the individuals who want to reduce collinearity as much as possible and deal at the construct level. We have tried for more of the latter than the former for a number of reasons. One reason is that we would like the project to produce as many generalizable truths as possible. Another stems from the dilemma between accuracy of prediction and accuracy of estimation, or the cross validation problem. That is, the more a prediction equation maximizes the accuracy of prediction in the sample, the more error it introduces into the estimation of the degree of accuracy in the population.

Project A is faced with the task of estimating several kinds of differential validity. It is reasonable to ask at the outset whether it is even possible, for a system of any multivariate complexity, to detect reasonable amounts of differential prediction with reasonable amounts of statistical power. The fewer parameters one must estimate, the greater the chances of being able to do that, which is a primary reason for examining the latent structure of predictors and criteria as carefully as possible.

Since we can draw a fairly reasonable picture of the population variance matrices for both predictors and criteria and thus provide a better starting point for Monte Carlo studies, one major research question we hope to answer is whether it is ever possible to estimate the parameters necessary for building a true classification algorithm. If it can't be done with a sample of 20 jobs and 500 cases per job, then perhaps the textbook discussions of the classification problem are a bit academic.

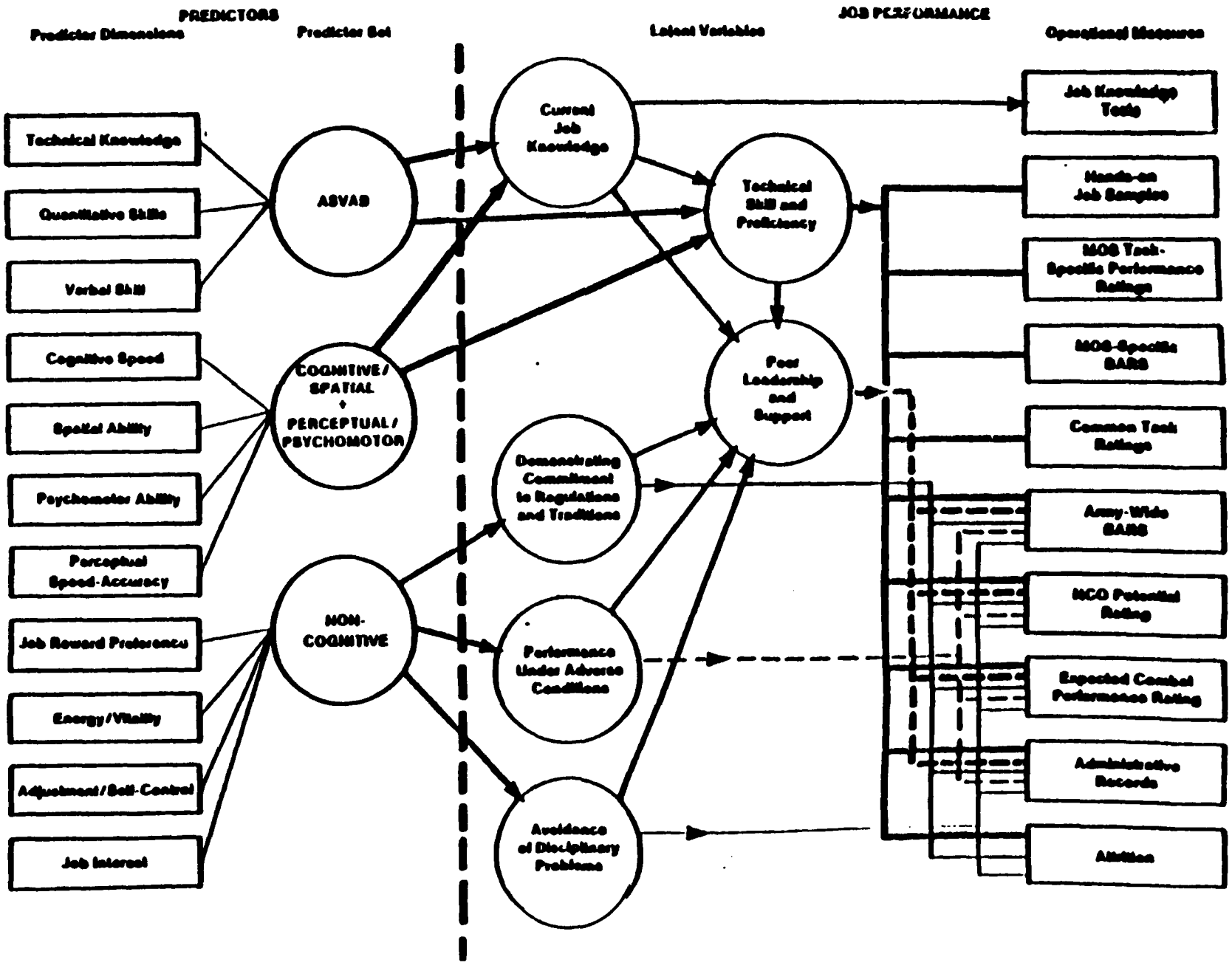
The Road to Constructs

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 the elusive and seldom seen college professor. Unfortunately, we were operating in a different job population and knew only that paper-and-pencil measures and rating measures would produce a lot of so-called method variance.

Given this initial disparity, we used both expert judgment and factor analytic results from the field tests to formulate a target model. A picture of that model is shown in Figure 3.

FIGURE 3

JOB PERFORMANCE — A PROPOSED STRUCTURAL MODEL



This picture is included only to show one stage in the almost continuous process of bootstrapping ourselves toward a more final conceptual description of the predictor/criterion space.

The target model was then subjected to what might be be described as a "quasi" confirmatory analysis using the concurrent validation sample. For the predictor scales that meant using the target to specify the number of factors for a full sample solution (i.e., all MOS combined). The predictor constructs and their associated component scales are shown in Table 5.

For the within MOS criterion matrixes we used confirmatory analyses and attempted to test alternative models. The alternative models were obtained by allowing the principal investigators to first look at the data, in the form of a series of principal component analyses, and to formulate a target matrix for a LISREL solution. Some clear alternative ideas emerged and these were compared in each MOS. After not too much cutting and fitting, we arrived at a single portrayal of the latent structure of performance that both fit the data in each job and seemed to make good sense. Obviously, the confirmatory analysis was not used in a strictly confirmatory way. This structure of job performance is portrayed in Table 6.

The model best confirmed by LISREL specified five "substantive" and "ratings" and "written test" method factors, that were orthogonal to the substantative factors and to each other. The first two substantive factors are based on the knowledge tests and the job sample measures. We have called these the core technical performance factor and the general (not so core) task performance factor. The technical factor reflects content which is central and largely specific to the MOS. The second factor encompasses content that tends to be common across several jobs and is less central to the core performance objectives. For this job population a significant part of the factor is represented in the common tasks, such as first aid, basic navigation, use of communication equipment, etc. However, it should be possible to make this distinction for virtually any job.

The remaining factors are based on the ratings, primarily those developed by the critical incident method, and the administrative/personnel records that were collected via self-report. Factor three encompassed the most scales and was the clearest in terms of its loading but the most heterogeneous appearing in terms of content. It appears to be a general effort and performance, performance under adverse conditions, peer leadership factor. In a spirit of wishful thinking, we had originally hoped to separate some of these elements, but either the lack of a distinct latent structure or the fallibility of the measures prevented it. Factor four is much more homogenous 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 five 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.

Table 5

Ability, temperament, and interest factors identified via analysis of the concurrent validity data on 9430 job incumbents. The tests and inventory scales from the trial battery which were used to form simple sum factor scores are listed under each factor title.

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)
- Pooled Mean Movement Time

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

Perceptual Accuracy Factor
- Short Term Memory test (Percent correct)
- Perceptual Speed & Accuracy test (Percent correct)
- 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 time)
- Number Memory test (Final decision time)

Simple Reaction Speed Factor
- Choice Reaction Time test (Decision time)
- Simple Reaction Time test (Decision time)

Simple Reaction Accuracy Factor
- Choice Reaction Time test (Percent correct)
- Simple Reaction Time test (Percent correct)

FROM NON-COGNITIVE INVENTORIES

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

Dependability Factor
- Conscientiousness scale
- Non-delinquency scale

Adjustment Factor
- Emotional Stability scale

Physical Condition Factor
- Physical Condition scale

Skilled Technical Interest Factor
- Clerical/Administrative
- Medical Services
- Leadership/Guidance
- Science/Chemical
- Data Processing
- Mathematics
- Electronic 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

Preference for Organizational & Co-worker Support
- Status
- Serving Others
- Organizational Support
- Ambition

Preference for Routine Work
- Routine

Preference for Job Autonomy
- Autonomy

Table 6

Performance factors representing the common latent structure across all jobs in the Project A sample. The criterion measures that comprise each factor are as indicated.

1) **Task Proficiency: Specific core technical skills:** The proficiency with which the individual performs the tasks which 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, standardized, and then added together for a total factor score. The factor score does not include any rating measures.

2) **Task Proficiency: General or common skills:** 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, etc. This factor represents proficiency on these general tasks.

- The same procedure (as for factor one) was used to compute the knowledge and hands-on general task scores, standardized within methods, and add the two standardized scores.

3) **Peer 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. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient?

- Five scales from the Army-wide BARS rating form (Technical Knowledge/Skill, Leadership, Effort, Self-development, and Maintaining Assigned Equipment), the expected combat performance scales, the job-specific BARS scales, the general performance rating, 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 (Following regulations, Self-Control, and Integrity) and two indices from the administrative records (number of disciplinary actions and promotion 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 two rating scales from the Army-wide BARS (Military Appearance and Physical Fitness).

In general, this solution fits the data from all MOS, seems reasonable and appropriate to Army management, and is not too far from our hypothesized structure, although we hoped to split factors two and three into a few more pieces.

Given these two pictures of the predictor domain and the performance space, we have begun exploring questions of differential validity across criterion components, differential validity across jobs, differential validity across subgroups for people, and overall classification efficiency under a variety of constraints.

Criterion Intercorrelations

As described in Wise, Campbell, Hanser, and McHenry (1986) five residual scores were created from the five criterion factors in the following manner. A paper-and-pencil "methods" factor score was created by first summing the two paper-and-pencil knowledge tests (job knowledge and training content knowledge scores) and then partialing out the variance due to the correlation of the total paper-and-pencil test score with all non-paper-and-pencil criterion measures (e.g., hands-on scores, rating scores, and administrative records scores). This residual was defined as the paper-and-pencil method score. This variable was in turn partialed from the Core Technical Proficiency criterion factor and from the General Task Proficiency factor creating two residual scores. A similar procedure was used to create a rating method factor score which was in turn partialed from the Effort/Leadership, Personal Discipline, and Physical Fitness/Military Bearing factors, thereby creating three more residual scores.

The five criterion factor scores, the five residual criterion scores, the single rating obtained from the overall performance rating scales, and the total score from the hands-on test were used to generate a 12 x 12 matrix of criterion intercorrelation for each MOS in Batch A. The averages of these correlations across MOS are shown in Table 7.

Remember that to create the residual scores the paper-and-pencil factor was partialed from the first two criterion factors and the rating method factor was partialed from the last three criterion factors. The intercorrelations of the 5 criterion factors are in the upper left quadrant, the intercorrelations among the 5 residual scores are in the lower right quadrant, and the cross correlations are in the upper right and lower left. Also remember that the first two factors contain items from both the knowledge tests and hands-on tests and the last three factors all contain both ratings and administrative measures.

Some noteworthy features of this 12 x 12 matrix are the following:

- The intercorrelations of the factor pairs which confound measurement method (e.g., 1 with 2 or 3 with 4) are higher, as expected, than factor pairs which do not confound method (e.g., 1 with 3 or 2 with 4). However, they are not so high that collapsing the

Table 7

Mean Intercorrelations among 12 Summary Criterion Measures
for the Batch A MOS

Criterion Summary Score	M3RAWCTP	M3RAWGSP	M3RAWELS	M3RAWMPD	M3RAWPFB	M3XPC11	M3XHTOTT	M3RESCTP	M3RESGSP	M3RESELS	M3RESMPD	M3RESPFB
M3RAWCTP: Core Tech Prof (raw)	1.000	0.531	0.280	0.190	0.032	0.243	0.737	0.880	0.380	0.465	0.225	0.040
M3RAWGSP: Gen Soldier Prof (raw)	0.531	1.000	0.268	0.163	0.041	0.206	0.722	0.386	0.891	0.451	0.192	0.047
M3RAWELS: Effort/Leadership (raw)	0.280	0.268	1.000	0.590	0.457	0.868	0.261	0.351	0.328	0.647	0.284	0.187
M3RAWMPD: Personal Discipline (raw)	0.190	0.163	0.590	1.000	0.335	0.650	0.150	0.256	0.226	0.436	0.894	0.194
M3RAWPFB: Fitness/Bearing (raw)	0.032	0.041	0.457	0.335	1.000	0.474	0.071	0.031	0.042	0.253	0.173	0.921
M3XPC11: Overall Perf Rating	0.243	0.206	0.868	0.650	0.474	1.000	0.203	0.309	0.261	0.444	0.333	0.192
M3XHTOTT: Hands-On Total	0.737	0.722	0.261	0.150	0.071	0.203	1.000	0.823	0.795	0.438	0.179	0.086
M3RESCTP: Core Tech Prof (resid)	0.880	0.386	0.351	0.256	0.031	0.309	0.823	1.000	0.440	0.453	0.246	-0.009
M3RESGSP: Gen Soldier Prof (resid)	0.380	0.891	0.328	0.226	0.042	0.261	0.795	0.440	1.000	0.432	0.212	0.007
M3RESELS: Effort/Leadership (resid)	0.465	0.451	0.647	0.436	0.253	0.444	0.438	0.453	0.432	1.000	0.477	0.277
M3RESMPD: Personal Discipline (resid)	0.225	0.192	0.284	0.894	0.173	0.333	0.179	0.246	0.212	0.477	1.000	0.200
M3RESPFB: Fitness/Bearing (resid)	0.040	0.047	0.187	0.194	0.921	0.192	0.086	-0.009	0.007	0.277	0.200	1.000

Table 8

Multiple Correlations[1] between
Criterion Scores
and Predictor Composite Scores Derived from Each Predictor Set Alone

Criterion Score	Predictor Composite: Composite Derived from ASVAB Factors (k = 4)[2]	Spatial Ability Factor (k = 1)	Composite Derived from Perceptual/ Psychomotor Computer Factors (k = 6)	Composite Derived from Biodata/ Temperament (ABLE) Factors (k = 4)	Composite Derived from Interest (AVOICE) Factors (k = 6)	Composite Derived from Job Reward Preference (JCB) Factors (k = 3)
Hands-On Total Score	.49	.46	.42	.20	.27	.22
Core Technical Proficiency (raw score)	.63	.57	.53	.25	.35	.27
Core Technical Proficiency (res score)[3]	.48	.40	.38	.21	.29	.21
General Soldiering Proficiency (raw score)	.66	.64	.58	.25	.35	.29
General Soldiering Proficiency (resid score)	.51	.50	.43	.21	.28	.22
Effort/ Leadership (raw score)	.33	.27	.27	.33	.25	.19
Effort/ Leadership (resid score)	.46	.42	.38	.31	.32	.25
Personal Discipline (raw score)	.20	.16	.15	.32	.15	.11
Personal Discipline (resid score)	.21	.18	.16	.28	.17	.10
Fitness/ Bearing (raw score)	.20	.11	.11	.36	.12	.11
Fitness/ Bearing (resid score)	.21	.12	.13	.34	.13	.10

[1]Multiple Rs are adjusted for shrinkage and corrected for range restriction, but are not corrected for criterion unreliability.
[2]k = the number of predictor factor scores used in computing the composite.
[3]Residual scores were formed by partialing a paper-and-pencil "method" construct from Core Technical and General Soldiering Proficiency and by partialing a rating "method" construct from Effort/Leadership, Personal Discipline and Fitness/Bearing.

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 inventories increases. This reversal becomes even more distinct for factors four and five. Notice that the interest scales are also a reasonably good predictor of task performance and do not predict factors three, four, and five as well as the temperament scales. The mixed nature of factor three is interesting and along with the confounding of method variance between the first two and the last three factors, it invites a consideration of residual scores.

For us at least, one of the most interesting aspects of the table is a comparison of the factor three raw score with the residualized factor three. As compared to the correlations with the raw score the correlations of the cognitive measures with the residual go up substantially and the correlations with the temperament composite go down slightly. The correlation of the interest composite with factor three also goes up when the rating method factor is partialed out. In general, interest in task content is more closely associated with task performance than with the more volitional nature of factors three, four, and five. These differences are not nearly so pronounced for the other two factors that involve ratings. We think this is because factor three includes the scales that in fact asked raters to assess the technical performance of the ratee. It is tempting to infer that raters are in fact influenced by the actual task competence of raters but that they also reflect differences in what might be termed dispositional or volitional behaviors of the kind predicted by personality/interest measures. Does the individual work hard, help others when they need it, keep going under adverse conditions, etc.? In our framework, these are both important components of performance and they are predicted by different things, but assessment via ratings cannot separate them very well. Perhaps it is also understandable why raters would have a difficult time separating them. It would require almost a mental partial correlation to do so.

Incremental Validities

The incremental validities for the new cognitive tests and new noncognitive tests over and above ASVAB alone for each of the performance factors can be obtained from the results presented in Table 9. While these comparisons are still at a rather general level and more analyses need to be done, one reasonable conclusion is that the new battery will provide the largest increments for the prediction of the "will do" aspects of performance. Also, we have not yet begun to consider what mix of ASVAB subtests and new cognitive tests might prove optimal for both selection and classification.

Table 9

Multiple Correlations[1] between
Criterion Scores
and Predictor Composite Scores Derived from Each Predictor Set Plus the ASVAB

Criterion Construct	Predictor Composite						
	Composite Derived from ASVAB Factors Alone	Composite Derived from ASVAB and Spatial Ability Factor	Composite Derived from ASVAB and Perceptual/ Psychomotor Computer Factors	Composite Derived from ASVAB and Biodata/ Temperament (ABLE) Factors	Composite Derived from ASVAB and Interest (AVOICE) Factors	Composite Derived from ASVAB and Job Reward Preference (JOB) Factors	Composite Derived from ASVAB and All Trial Battery (TB) Factors
Hands-On Total Score	.49	.51	.51	.49	.50	.49	.53
Core Technical Proficiency (raw score)	.63	.65	.64	.64	.65	.64	.67
Core Technical Proficiency (res score)[2]	.48	.49	.49	.49	.50	.48	.52
General Soldiering Proficiency (raw score)	.66	.69	.68	.67	.67	.67	.71
General Soldiering Proficiency (resid score)	.51	.53	.52	.51	.52	.51	.55
Effort/ Leadership (raw score)	.33	.34	.34	.43	.37	.35	.45
Effort/ Leadership (resid score)	.46	.47	.47	.51	.49	.47	.53
Personal Discipline (raw score)	.20	.20	.20	.37	.23	.23	.38
Personal Discipline (resid score)	.21	.21	.22	.35	.24	.23	.36
Fitness/ Bearing (raw score)	.20	.21	.21	.41	.24	.22	.42
Fitness/ Bearing (resid score)	.21	.23	.24	.40	.25	.23	.41

[1]Multiple Rs are adjusted for shrinkage and corrected for range restriction, but are not corrected for criterion unreliability.
[2]Residual scores were formed by partialing a paper-and-pencil "method" construct from Core Technical and General Soldiering Proficiency and by partialing a rating "method" construct from Effort/Leadership, Personal Discipline and Fitness/Bearing.

Table 10

Results of Stepwise Regressions within Each Predictor Domain
for the Four Army-wide Performance Constructs
across All 9 Batch A MOS

Predictor Constructs	Criterion Construct				
	General Soldiering (raw score)	Effort and Leadership (resid score)	Effort and Leadership (raw score)	Personal Discipline (raw score)	Phys Fitness/ Mil Bearing (raw score)
ASVAB FACTORS					
Verbal	0.10	0.03	-0.07	-0.03	-0.11
Quantative	0.20	0.08	0.03	0.07	0.03
Technical	0.26	0.21	0.21	0.06	-0.05
Speed	0.03	0.07	0.09	0.04	0.10
ADJ, UNCORR R	0.461	0.280	0.206	0.106	0.161
SPATIAL					
Overall Spatial	0.47	0.25	0.14	0.07	-0.05
UNCORRECTED R	0.466	0.253	0.142	0.068	0.047
COMPUTER					
Complex Perc Speed	-0.09	-0.06	-0.07	.	.
Complex Perc Accy	0.19	0.07	0.09	0.05	.
Number Speed/Accy	-0.14	-0.06	-0.09	-0.03	.
Psychomotor	-0.19	-0.08	-0.10	.	.
Simp Reaction Accy	0.04	.	.	.	-0.06
Simp Reaction Speed	.	.	.	.	-0.07
ADJ, UNCORR R	0.363	0.149	0.208	0.032	0.071
TEMPERAMENT					
Adjustment	0.09	0.04	0.03	0.03	.
Dependability	0.04	.	0.06	0.30	0.12
Surgency	0.04	0.23	0.25	.	0.12
Phys Condition	-0.06	.	.	-0.06	0.24
ADJ, UNCORR R	0.129	0.255	0.303	0.303	0.356
INTERESTS					
Combat	0.24	0.20	0.17	.	0.04
Machines	.	.	.	-0.04	-0.06
Audiovisual	.	.	-0.04	.	.
Technical	.	0.06	0.08	0.09	0.14
Food Service	-0.10	-0.16	-0.12	-0.06	-0.05
Protective Svc	-0.06	.	.	-0.09	.
ADJ, UNCORR R	0.229	0.235	0.199	0.078	0.119
JOB VALUES					
Support	.	0.03	0.05	0.05	0.10
Autonomy	0.05	0.07	0.03	-0.06	-0.05
Routine	-0.11	-0.12	-0.09	-0.03	-0.02
ADJ, UNCORR R	0.123	0.150	0.112	0.063	0.097

Table 11

Results of Stepwise Regressions within Each Predictor Domain
for MOS-Specific Core Technical Proficiency
for Each of the 9 Batch A MOS

	MOS								
Predictor Constructs	11B	13B	19E	31C	63B	64C	71L	91A	95B
ASVAB FACTORS									
Verbal	0.20	.	0.13	0.19	.	.	0.16	0.25	0.11
Quantative	0.14	0.09	0.15	0.14	.	0.14	0.38	0.12	0.16
Technical	0.23	0.23	0.27	0.23	0.55	0.34	-0.11	0.19	0.11
Speed	0.10	.	.	0.11	.	.	0.08	0.17	0.09
ADJ, UNCORR R	0.503	0.254	0.452	0.427	0.538	0.413	0.441	0.456	0.282
SPATIAL									
Overall Spatial	0.48	0.33	0.43	0.32	0.41	0.37	0.41	0.38	0.28
UNCORRECTED R	0.475	0.334	0.432	0.315	0.412	0.366	0.411	0.380	0.275
COMPUTER									
Comp Perc Speed	-0.25	-0.10	.	.	-0.08	-0.14	.	.	.
Comp Perc Accy	0.29	0.11	0.16	0.13	.	0.19	0.27	0.09	0.13
Number Speed/Accy	-0.11	-0.11	-0.20	-0.25	-0.08	-0.07	-0.22	-0.20	-0.19
Psychomotor	-0.13	-0.17	-0.11	-0.09	-0.20	-0.10	.	-0.15	-0.09
Simp Reaction Accy	.	.	0.12	.	0.08	0.07	.	0.08	.
Simp Reaction Speed	.	.	.	.	.	.	.	.	.
ADJ, UNCORR R	0.406	0.257	0.343	0.253	0.242	0.269	0.326	0.261	0.228
TEMPERAMENT									
Adjustment	.	0.12	0.14	.	0.10	.	.	0.10	0.08
Dependability	.	.	0.08	0.10		.	0.10	0.19	0.12
Surgency	0.19	.	.	.	0.09	.	0.14	.	.
Phys Condition	.	.	-0.13	.	-0.12	.	-0.10	-0.15	.
ADJ, UNCORR R	0.143	0.000	0.129	0.000	0.119	0.000	0.176	0.211	0.114
INTERESTS									
Combat	0.25	0.25	0.26	.	0.11	0.09	0.12	0.18	.
Machines	.	0.10	.	0.13	0.38	0.09	-0.23	.	.
Audiovisual	.	.	.	.	-0.11	.	.	.	-0.08
Technical	0.08	.	.	0.10	.	.	0.19	.	.
Food Service	-0.22	-0.16	-0.11	.	-0.10	-0.12	-0.07	.	-0.06
Protective Svc	-0.11	-0.10	.	.	-0.14	.	.	.	.
ADJ, UNCORR R	0.276	0.255	0.218	0.000	0.441	0.135	0.160	0.039	0.000
JOB VALUES									
Support	.	.	.	.	.	.	.	0.14	.
Autonomy	0.08	0.17	.	.	0.14	0.11	.	.	.
Routine	-0.15	-0.14	-0.21	.	-0.10	-0.07	-0.12	.	-0.08
ADJ, UNCORR R	0.141	0.201	0.166	0.000	0.133	0.080	0.038	0.058	0.000

Table 12

Results of Stepwise Regressions
for the Four Army-Wide Performance Constructs
across All 9 Batch A MOS

Predictor Constructs	Criterion Construct				
	General Soldiering (raw score)	Effort and Leadership (resid score)	Effort and Leadership (raw score)	Personal Discipline (raw score)	Phys Fitness/ Mil Bearing (raw score)
ASVAB FACTORS					
Verbal	0.09	0.03	-0.06	.	-0.10
Quantitative	0.09	0.04	.	0.05	.
Technical	0.12	0.11	0.15	0.07	-0.03
Speed	.	0.04	0.06	0.03	0.08
SPATIAL					
Overall Spatial	0.25	0.13	.	.	.
COMPUTER					
Complex Perc Speed	.	.	-0.05	.	.
Complex Perc Accy	0.08	.	0.04	.	.
Number Speed/Accy	-0.02	.	.	0.03	.
Psychomotor	-0.04	.	-0.02	.	.
Simp Reaction Accy	.	.	.	.	-0.04
Simp Reaction Speed	-0.03	.	.	.	-0.05
TEMPERAMENT					
Adjustment	.	.	.	.	.
Dependability	0.11	0.06	0.11	0.30	0.09
Surgency	-0.04	0.15	0.20	0.03	0.14
Phys Condition	.	0.03	.	-0.05	0.22
INTERESTS					
Combat	0.13	0.11	0.10	.	0.04
Machines	.	.	.	.	-0.05
Audiovisual	.	-0.02	-0.04	-0.03	0.04
Technical	.	.	.	.	.
Food Service	-0.04	-0.08	-0.06	-0.04	.
Protective Svc	.	0.03	.	-0.03	-0.05
JOB VALUES					
Support	.	.	.	.	.
Autonomy	.	.	.	-0.05	-0.04
Routine	-0.03	-0.04	-0.03	.	.
ADJUSTED, UNCORRECTED R	0.540	0.392	0.366	0.317	0.385

Table 13

Results of Stepwise Regressions
for MOS-Specific Core Technical Proficiency
for Each of the 9 Batch A MOS

	MOS								
Predictor Constructs	11B	13B	19E	31C	63B	64C	71L	91A	95B
ASVAB FACTORS									
Verbal	0.17	.	0.10	0.21	.	.	0.08	0.26	0.13
Quantitative	0.09	.	.	0.30	.	.	0.27	.	.
Technical	0.10	.	0.16	.	0.35	0.30	-0.13	0.12	.
Speed	.	.	.	.	.	-0.07	.	0.13	.
SPATIAL									
Overall Spatial	0.20	0.25	0.19	.	0.14	0.16	0.25	0.23	0.22
COMPUTER									
Complex Perc Speed	-0.18	.	.	.	.	-0.12	.	.	.
Complex Perc Accy	0.13	.	0.09	-0.10	.	0.14	0.15	.	0.09
Number Speed/Accy	.	.	-0.09	.	.	.	.	.	-0.11
Psychomotor	.	.	.	.	.	.	.	.	.
Simp Reaction Accy	.	.	0.07	.	.	.	.	.	.
Simp Reaction Speed	.	-0.10	.	.	-0.11	.	.	.	.
TEMPERAMENT									
Adjustment	-0.08	.	.	-0.09	.	.	.	.	.
Dependability	0.12	.	0.10	0.15	0.13	0.07	0.11	0.22	0.12
Surgency	.	.	.	.	.	.	.	.	.
Phys Condition	.	.	-0.09	.	-0.06	.	.	-0.13	.
INTERESTS									
Combat	0.15	0.21	0.17	.	.	.	.	0.16	.
Machines	.	.	.	0.21	0.32	.	-0.16	.	.
Audiovisual	.	.	.	.	-0.14	.	.	-0.09	-0.13
Technical	.	.	.	.	.	.	0.12	.	.
Food Service	-0.07	.	.	.	.	.	.	.	.
Protective Svc	.	-0.08	.	.	-0.08	.	.	.	.
JOB PREFERENCES									
Support	.	.	.	.	.	0.09	.	0.12	0.09
Autonomy	.	0.09	.	-0.11	.	.	.	.	.
Routine	-0.06	-0.11	.	.	.	.	.	0.07	.
ADJUSTED, UNCORRECTED R	0.560	0.305	0.464	0.352	0.591	0.401	0.481	0.507	0.294

Table 14

Correlations between the Predictor Constructs
and the Army-Wide Criterion Constructs
Combined across Batch A MOS
(Corrected for Range Restriction)

Predictor Construct	Criterion Construct				
	General Soldiering (raw score)	Effort and Leadership (resid score)	Effort and Leadership (raw score)	Personal Discipline (raw score)	Phys Fitness/ Mil Bearing (raw score)
ASVAB Factor: Technical	0.55	0.39	0.28	0.12	-0.08
ASVAB Factor: Verbal	0.52	0.35	0.20	0.10	-0.07
ASVAB Factor: Quantative	0.54	0.36	0.23	0.14	-0.01
ASVAB Factor: Speed	0.37	0.29	0.21	0.11	0.07
Cognitive Construct: Overall Spatial	0.59	0.38	0.24	0.11	-0.03
Computer Construct: Complex Perc Speed	-0.21	-0.17	-0.13	-0.03	-0.04
Computer Construct: Complex Perc Accy	0.30	0.18	0.12	0.08	-0.01
Computer Construct: Number Speed/Accy	-0.44	-0.31	-0.21	-0.09	-0.01
Computer Construct: Psychomotor	-0.40	-0.27	-0.20	-0.04	-0.01
Computer Construct: Simp Reaction Accy	0.18	0.09	0.05	0.05	-0.05
Computer Construct: Simp Reaction Speed	-0.19	-0.13	-0.08	-0.01	-0.06
ABLE Construct: Adjustment	0.18	0.22	0.23	0.13	0.17
ABLE Construct: Physical Condition	-0.03	0.09	0.10	-0.02	0.30
ABLE Construct: Dependability	0.09	0.15	0.21	0.30	0.22
ABLE Construct: Surgency	0.16	0.30	0.33	0.20	0.27
AVOICE Construct: Audiovisual Arts	0.02	0.02	0.01	0.00	0.07
AVOICE Construct: Combat Related	0.23	0.22	0.19	-0.00	0.03
AVOICE Construct: Food Service	-0.12	-0.14	-0.11	-0.06	-0.00
AVOICE Construct: Structural/Machines	0.06	0.06	0.06	-0.05	-0.01
AVOICE Construct: Protective Services	-0.04	0.03	0.04	-0.04	0.02
AVOICE Construct: Skilled Technical	0.04	0.07	0.06	0.05	0.11
JOB Construct: Autonomy	0.13	0.15	0.09	-0.02	-0.02
JOB Construct: Routine	-0.21	-0.20	-0.15	-0.06	-0.04
JOB Construct: Org & Coworker Support	0.09	0.11	0.10	0.05	0.09

Table 15

Correlations between the Predictor Constructs
and Core Technical Proficiency
(Corrected for Range Restriction)

	MOS								
Predictor Construct	11B	13B	19E	31C	63B	64C	71L	91A	95B
ASVAB Factor: Technical	0.60	0.36	0.56	0.59	0.69	0.55	0.37	0.61	0.51
ASVAB Factor: Verbal	0.63	0.33	0.49	0.67	0.50	0.44	0.56	0.71	0.59
ASVAB Factor: Quantative	0.60	0.32	0.49	0.67	0.45	0.46	0.63	0.64	0.59
ASVAB Factor: Speed	0.48	0.25	0.28	0.57	0.29	0.27	0.52	0.56	0.47
Cognitive Construct: Overall Spatial	0.63	0.41	0.55	0.58	0.56	0.51	0.57	0.64	0.56
Computer Construct: Complex Perc Speed	-0.33	-0.15	-0.17	-0.25	-0.24	-0.25	-0.11	-0.28	-0.20
Computer Construct: Complex Perc Accy	0.35	0.24	0.32	0.22	0.16	0.28	0.40	0.25	0.26
Computer Construct: Number Speed/Accy	-0.48	-0.30	-0.42	-0.62	-0.37	-0.38	-0.50	-0.57	-0.53
Computer Construct: Psychomotor	-0.43	-0.30	-0.36	-0.34	-0.36	-0.34	-0.26	-0.44	-0.32
Computer Construct: Simp Reaction Accy	0.17	0.11	0.26	0.17	0.14	0.19	0.27	0.16	0.20
Computer Construct: Simp Reaction Speed	-0.17	-0.19	-0.15	-0.10	-0.23	-0.19	-0.11	-0.21	-0.23
ABLE Construct: Adjustment	0.26	0.13	0.18	0.06	0.21	0.07	0.20	0.12	0.27
ABLE Construct: Physical Condition	0.06	-0.04	-0.09	-0.18	-0.13	-0.07	-0.12	-0.09	-0.13
ABLE Construct: Dependability	0.16	0.01	0.09	0.04	0.00	0.01	0.21	0.18	0.24
ABLE Construct: Surgency	0.31	0.06	0.16	0.14	0.20	0.09	0.27	0.22	0.25
AVOICE Construct: Audiovisual Arts	0.04	-0.05	-0.01	0.20	-0.14	-0.00	0.19	0.13	-0.14
AVOICE Construct: Combat Related	0.23	0.21	0.31	0.08	0.31	0.24	0.02	0.22	0.03
AVOICE Construct: Food Service	-0.30	-0.14	-0.14	0.01	-0.20	-0.14	-0.03	-0.09	-0.19
AVOICE Construct: Structural/Machines	-0.12	0.09	0.06	0.05	0.41	0.16	-0.19	0.01	-0.19
AVOICE Construct: Protective Services	-0.05	-0.08	-0.04	-0.01	-0.10	-0.05	0.01	-0.13	-0.16
AVOICE Construct: Skilled Technical	0.07	-0.03	0.09	0.12	-0.08	0.00	0.17	0.00	-0.03
JOB Construct: Autonomy	0.21	0.22	0.09	0.22	0.25	0.21	0.21	0.23	0.09
JOB Construct: Routine	-0.27	-0.18	-0.27	-0.19	-0.21	-0.20	-0.19	-0.22	-0.30
JOB Construct: Org & Coworker Support	0.14	0.13	0.05	-0.02	0.06	0.14	0.20	0.18	-0.01

Moving on from this point our future validity analyses will be concerned with:

- More precise estimates of validity generalization across jobs as a function of criterion content and predictor battery composition.
- Estimation of differential prediction across race and gender groups as a function of criterion content and predictor battery composition.
- Estimation of overall selection validity (against a criterion composite) as a function of criterion component weights and predictor battery composition.
- Estimation of classification efficiency.

REFERENCES

Campbell, J. P. (1986). When the textbook goes operational. Paper presented at the 94th Annual Convention of the American Psychological Association, Washington, D.C.

Dunnette, M. D. (1963). A modified model for selection research. Journal of Applied Psychology, 47, 317-323.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology, 24, 419-434.

Wallace, S. R. (1965). Criteria for what? American Psychologist, 20, 411-412.

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

A LATENT STRUCTURE MODEL OF JOB PERFORMANCE FACTORS: APPENDIX*

Jeffrey J. McHenry
Lauress L. Wise
American Institutes for Research

John P. Campbell
Human Resources Research Organization

Lawrence M. Hanser
U.S. Army Research Institute

Presented at a Data Analysis Workshop of the
Committee on Performance of Military Personnel

Baltimore

December 1986

* This Appendix supplements the paper,, "A Latent Structure Model of Job Performance Factors," first presented at the Convention of the American Psychological Association in August 1986, and available in ARI Research Note 813704.

ASSUMPTIONS ABOUT JOB PERFORMANCE IN ENTRY-LEVEL ENLISTED MOS

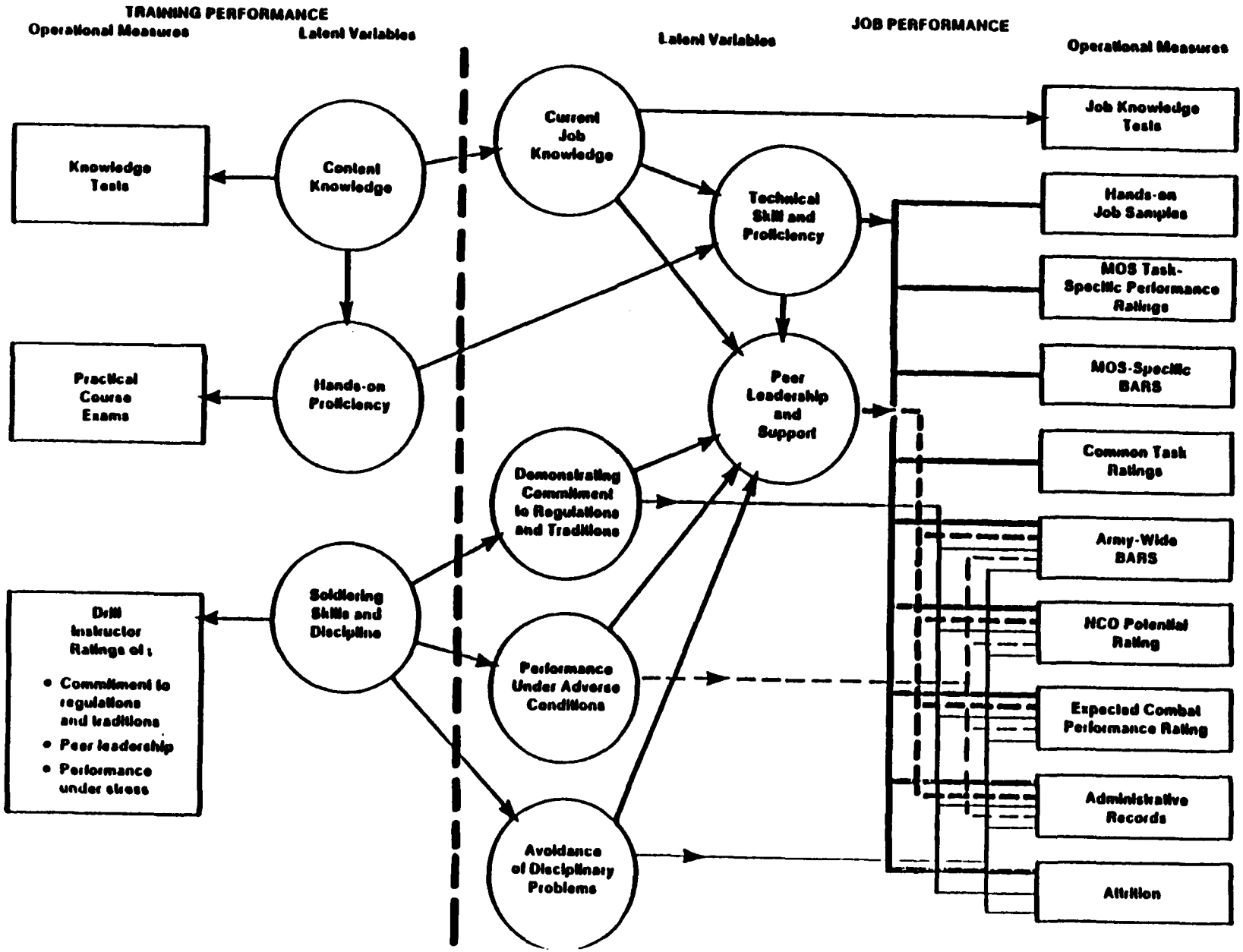
- Performance is not one thing. It is genuinely multidimensional. Consequently, no one measure can be identified as the measure of performance. There is no ultimate criterion. For example, a job simulation, no matter how elaborate, reliable, and valid, is not an ultimate measure. All measures have their measurement flaws, and all measures are constrained to some portion of the job performance domain.

- For the population of first-tour enlisted MOS, it makes sense to talk about job performance factors that are defined the same way across jobs (i.e., Army-wide factors) and performance factors that are specific to a particular job. The job-specific factors are the latent variables that capture the differences in task content that require different knowledges, skills, and abilities (KSAs). The Army-wide factors reflect tasks and other components of job performance that do not require different KSAs.

- While the content of job-specific factors and the degree of required performance on any particular factor may differ across jobs, the latent structure or basic form of performance is the same for all skilled entry level jobs in the military.

- To form criterion composites that represent an overall performance score for test validation purposes, or some other purpose that requires a single score, the individual latent variables must be measured, scored, weighted, and combined in some fashion. When forming criterion composites for different jobs, it is not the form of the latent variables that changes, but their relative weight. Also, it is a value judgment of the organization as to how much a particular performance component should be weighted for a particular measurement purpose.

PRELIMINARY MODEL OF ENLISTED JOB PERFORMANCE



MEASUREMENT METHODS

- Army-Wide BARS
- MOS-Specific BARS
- Combat Performance Prediction Scales
- Administrative Measures
- School Knowledge Test
- Job Knowledge Test
- Hands-On Performance Test

ARMY-WIDE BEHAVIORALLY-ANCHORED RATING SCALES (BARS) SUMMARY FACTORS

- Effort and Leadership
- Personal Discipline
- Physical Fitness and Military Bearing
- Overall Effectiveness

MOS-SPECIFIC BEHAVIORALLY-ANCHORED RATING SCALES (BARS) SUMMARY FACTORS

- Core Responsibilities
- Other Responsibilities

COMBAT PERFORMANCE PREDICTION SCALES

SUMMARY FACTORS

- Performing Well under Adverse Conditions
- Avoiding Mistakes

ADMINISTRATIVE MEASURES

SUMMARY "FACTORS"

- Letters and Certificates
- Physical Readiness Test Score
- M16 Qualification Score
- Articles 15/Flag Actions
- Promotion Rate Deviation Score

SAMPLE FUNCTIONAL DUTY CLUSTER DEFINITIONS

First Aid

Consists of items whose primary purpose is to indicate knowledge about how to sustain life, prevent health complications caused by trauma or environmentally induced illness, including the practice of personal hygiene. Includes all related diagnostic, transportation, and treatment items except those items normally performed in a patient care facility. Includes items related to safety and safety hazards.

Navigate

Consists of items whose primary purpose is to indicate knowledge about how to plan or execute movement between points over unknown terrain either cross-country or using road networks, or identify the location of objects. Includes all means of determining direction, distances, and locations using maps of all types, overlays, compasses, terrain, celestial objects, and field expedients.

NBC

Consists of items whose primary purpose is to indicate knowledge about performance when nuclear, biological, or chemical contaminants and threats are present, planned, detected, or expected. Includes maintenance and operation of clothing, gear, and equipment whose primary purpose is to counter, protect, or detect NBC threats. Includes NBC markers. Does not include first aid treatment of contamination.

Weapons

Consists of items whose primary purpose is to indicate knowledge about maintenance, preparation, and firing of small arms. Small arms are defined as sized weapons, including automatic weapons, up to and including caliber .60 and shotguns. Includes ancillary sighting systems and techniques, stands and mounts, zeroing and techniques of fire. Excludes firing from aircraft and vehicles where the weapon is fired by electrical/hydraulic aiming/firing systems and sighting systems that are part of the aircraft/vehicle and not part of the weapon.

FUNCTIONAL DUTY CLUSTERS BY MOS

Cluster Number and Name	11B			13B			19E			31C			63B			64C			71L			91A			95B		
	MO	JK	SK	MO	JK	SK	MO	JK	SK	MO	JK	SK	MO	JK	SK	MO	JK	SK	MO	JK	SK	MO	JK	SK	MO	JK	SK
1. First Aid	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2. Navigate		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3. NBC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X
4. Weapons	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X		X	X	X
5. Field Techniques	X	X	X	X	X	X		X	X		X	X		X	X		X	X		X	X		X	X	X	X	X
6. Communication	X	X	X	X	X	X	X	X	X	X	X				X			X			X				X	X	X
7. ID Target		X	X		X			X	X		X	X					X						X			X	X
8. Customs and Laws			X		X	X		X	X		X	X	X	X	X		X	X		X	X			X			X
9. Antitank/Antiair Weapons		X	X			X						X						X			X						X
11. Drive (Operate and Maintain)	X		X			X				X	X	X			X	X	X	X			X		X	X	X	X	X
14. Prepare/Operate/Maintain Howitzer and Ammunition				X	X	X																					
15. Operate Howitzer Sight/Alignment Device				X	X	X																					
16. Preventive Maintenance															X			X									X
17. Tank Operations							X	X	X																		
18. Tank Gunnery							X	X	X																		
20. Generators										X	X																
21. TTY Station and Net Operators										X	X																
22. Maintain TTY Electronic Equipment											X	X															
23. Operate TTY Electronic Equipment										X	X																
24. Install TTY Electronic Equipment										X	X																
25. Electrical System													X	X	X												
26. Brake/Steering/Suspension System													X	X	X												
27. Vehicle Operation and Recovery													X	X	X			X									X
28. Fuel/Cooling/Lubricating													X	X	X												
29. Forms/Files Management																			X	X	X						
30. Supervision/Coordination																				X	X						
31. Correspondence																			X	X	X						
32. Classified Material																			X	X	X						
33. Clinic/Ward Treatment and Care																						X	X	X			
34. Clinic/Ward Housekeeping																						X	X	X			
35. Clinic/Ward Management																							X	X			
36. General Medical Knowledge																								X			
37. Responding To Alarms																										X	X
38. Conduct MP Procedures																										X	X
39. Patrol Duties																									X	X	X
93. Power Train and Clutch													X	X													

HANDS-ON, JOB KNOWLEDGE, AND SCHOOL KNOWLEDGE TESTS

SUMMARY FACTORS

- Core Technical (MOS-specific)
- Communications
- Vehicle Operation and Maintenance
- General Soldiering
- Identifying Target and Threat Vehicles and Aircraft
- Safety and Survival

REASONS FOR IDENTIFYING A REDUCED SET OF PERFORMANCE CONSTRUCTS

FROM THE WITHIN-METHOD PERFORMANCE FACTORS

- Desire for performance indices that incorporate information from multiple measurement methods
- Parsimony (high correlations between many of the factors)
- Reliability
- Construct weighting

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 11B: INFANTRY

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	Overall Rating	4.60	0.90	.	90	74	68	77	85	75	65	23	12	17	-35	36	26	14	4	35	25	11	10	33	19	18	12	14
2	Eff/Ldr Rating	4.41	0.82	90	.	74	65	80	88	80	67	24	8	13	-30	36	30	12	5	36	27	10	13	33	20	20	9	17
3	Discipline Rtng	4.50	0.87	74	74	.	49	55	71	63	66	13	3	7	-39	31	16	10	3	30	22	6	8	24	13	13	5	13
4	Fitness Rating	4.86	0.89	68	65	49	.	59	66	52	45	17	27	9	-24	22	10	9	-1	10	10	-2	-4	13	6	6	1	1
5	Job-Spec Tech	32.98	4.58	77	80	55	59	.	86	75	58	23	15	17	-20	22	27	15	5	35	22	12	10	36	21	23	9	16
6	Job-Spec Other	22.67	3.66	85	88	71	66	86	.	80	67	25	8	14	-28	32	23	10	6	35	26	12	12	33	17	22	11	17
7	Combat Exmplry	9.02	1.49	75	80	63	52	75	80	.	75	24	8	13	-31	29	28	12	7	37	25	9	16	34	22	23	9	19
8	Combat Problems	10.03	1.64	65	67	66	45	58	67	75	.	14	8	6	-33	27	20	7	-1	36	24	9	15	31	21	18	8	14
9	Awards & Certs	3.33	2.18	23	24	13	17	23	25	24	14	.	15	20	-2	4	13	6	-1	14	15	-0	13	9	9	5	4	12
10	Phys. Readiness	273.44	28.00	12	8	3	27	15	8	8	8	15	.	11	2	-6	1	-7	-9	0	5	-7	-0	8	-2	-1	-4	-6
11	M16 Qualific.	2.74	0.57	17	13	7	9	17	14	13	6	20	11	.	1	1	13	6	-0	10	2	3	0	14	10	5	3	6
12	Articles 15	0.39	0.85	-35	-30	-39	-24	-20	-28	-31	-33	-2	2	1	.	-45	-10	-1	-6	-10	-9	-6	-6	-10	-1	-9	0	-5
13	Promotion Rate	0.03	0.68	36	36	31	22	22	32	29	27	4	-6	1	-45	.	16	7	7	19	17	12	10	18	14	12	11	17
14	MOS Basic	50.50	10.06	26	30	16	10	27	23	28	20	13	1	13	-10	16	.	15	6	44	30	13	27	40	24	20	16	30
15	MOS Safety	22.67	3.41	14	12	10	9	15	10	12	7	6	-7	6	-1	7	15	.	2	16	8	1	8	16	7	3	3	4
16	MOS Comm	13.15	1.53	4	5	3	-1	5	6	7	-1	-1	-9	-0	-6	7	6	2	.	4	6	-1	-3	0	4	6	2	-1
17	JK Basic	50.93	9.71	35	36	30	10	35	35	37	36	14	0	10	-10	19	44	16	4	.	66	40	42	65	50	40	30	35
18	JK Safety	20.02	4.31	25	27	22	10	22	26	25	24	15	5	2	-9	17	30	8	6	68	.	23	26	47	41	32	25	20
19	JK Comm	4.37	1.47	11	10	6	-2	12	12	9	9	-0	-7	3	-6	12	13	1	-1	40	23	.	16	26	25	19	14	18
20	JK Identify	8.25	2.24	10	13	8	-4	10	12	16	15	13	-0	0	-6	10	27	8	-3	42	26	16	.	31	24	18	16	37
21	SK Basic	72.87	14.89	33	33	24	13	36	33	34	31	9	8	14	-10	18	40	16	0	65	47	26	31	.	63	60	44	43
22	SK Safety	9.51	2.12	19	20	13	6	21	17	22	21	9	-2	10	-1	14	24	7	4	50	41	25	24	63	.	45	34	26
23	SK Comm	5.68	1.67	18	20	13	6	23	22	23	18	5	-1	5	-9	12	20	3	6	40	32	19	18	60	45	.	40	31
24	SK Vehicle	0.78	0.42	12	9	5	1	9	11	9	8	4	-4	3	0	11	16	3	2	30	25	14	16	44	34	40	.	21
25	SK Identify	2.80	1.16	14	17	13	1	16	17	19	14	12	-6	6	-5	17	30	4	-1	35	20	18	37	43	26	31	21	.

N= 503

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 13B: CANNON CREWMAN

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	Overall Rating	4.59	0.79	.	86	71	61	62	72	73	61	11	10	5	-25	30	20	19	17	6	26	18	14	8	3	24	15	12	8	9
2	Eff/Ldr Rating	4.43	0.76	86	.	75	62	65	74	78	61	14	6	1	-23	25	27	25	14	9	32	20	15	11	5	30	20	15	5	13
3	Discipline Rtng	4.61	0.78	71	75	.	51	53	60	63	60	-0	-4	-1	-20	26	12	9	12	4	22	16	15	4	3	18	14	14	6	16
4	Fitness Rating	4.95	0.82	61	62	51	.	47	53	51	39	7	23	-1	-25	16	8	4	0	3	5	-1	-1	1	-2	4	-4	-1	-4	-6
5	Job-Spec Tech	23.59	3.55	62	65	53	47	.	80	60	39	11	10	1	-2	10	35	18	9	-1	25	10	10	17	8	24	8	12	6	4
6	Job-Spec Other	23.90	3.08	72	74	60	53	80	.	66	49	6	5	-4	-9	18	25	18	8	1	29	18	15	13	6	26	14	16	4	8
7	Combat Exmplry	9.00	1.44	73	78	63	51	60	66	.	63	14	10	3	-15	23	20	23	13	3	22	16	13	6	8	23	12	7	-1	1
8	Combat Problems	9.92	1.56	61	61	60	39	39	49	63	.	8	7	-3	-16	26	14	16	6	12	19	17	10	14	8	15	14	9	5	8
9	Awards & Certs	2.58	1.82	11	14	-0	7	11	6	14	8	.	12	18	0	8	15	19	15	-1	11	10	6	5	8	11	6	5	8	3
10	Phys. Readiness	261.74	32.70	10	6	-4	23	10	5	10	7	12	.	11	-3	-2	7	2	-7	8	-8	-8	-10	5	4	-0	-8	-10	-12	-15
11	M16 Qualific.	2.25	0.69	5	1	-1	-1	1	-4	3	-3	18	11	.	6	1	7	8	12	-3	-4	-5	-6	7	-3	-3	-7	0	3	-3
12	Articles 15	0.46	1.03	-25	-23	-20	-25	-2	-9	-15	-16	0	-3	6	.	-31	-0	-4	-5	-5	-7	-10	-12	-7	1	-5	-6	-2	-5	-3
13	Promotion Rate	0.01	0.63	30	25	26	16	10	18	23	26	8	-2	1	-31	.	6	10	10	3	10	6	5	5	-1	2	5	-2	10	7
14	MQ Tech.	50.71	9.94	20	27	12	8	35	25	20	14	15	7	7	-0	6	.	47	20	11	33	13	7	10	12	36	18	20	11	9
15	MQ Basic	48.50	13.00	19	25	9	4	18	18	23	16	19	2	8	-4	10	47	.	21	8	42	38	20	9	15	40	25	17	15	9
16	MQ Safety	40.16	6.23	17	14	12	0	9	8	13	8	15	-7	12	-5	10	20	21	.	11	24	14	11	9	3	25	20	16	11	24
17	MQ Comm	10.60	1.59	6	9	4	3	-1	1	3	12	-1	8	-3	-5	3	11	8	11	.	1	1	-2	6	5	7	5	-1	1	3
18	JK Tech.	50.67	9.94	26	32	22	5	25	29	22	19	11	-8	-4	-7	10	33	42	24	1	.	58	54	21	20	64	52	41	37	25
19	JK Basic	31.91	5.78	18	20	16	-1	10	18	16	17	10	-8	-5	-10	6	13	38	14	1	58	.	55	14	23	52	49	38	35	27
20	JK Safety	23.58	4.43	14	15	15	-1	10	15	13	10	6	-10	-6	-12	5	7	20	11	-2	54	55	.	10	21	41	38	35	26	27
21	JK Comm	1.12	0.68	8	11	4	1	17	13	6	14	5	5	7	-7	5	10	9	9	6	21	14	10	.	13	19	13	16	14	11
22	JK Identify	7.12	2.25	3	5	3	-2	8	6	8	8	8	4	-3	1	-1	12	15	3	5	20	23	21	13	.	20	21	25	10	8
23	SK Tech.	50.82	9.84	24	30	18	4	24	26	23	15	11	-0	-3	-5	2	36	40	25	7	64	52	41	19	20	.	63	47	38	40
24	SK Basic	23.17	5.27	15	20	14	-4	8	14	12	14	6	-8	-7	-6	5	18	25	20	5	52	49	36	13	21	63	.	51	40	52
25	SK Safety	8.44	2.12	12	15	14	-1	12	16	7	9	5	-10	0	-2	-2	20	17	18	-1	41	38	25	16	25	47	51	.	28	36
26	SK Comm	3.55	1.21	8	5	6	-4	6	4	-1	5	8	-12	3	-5	10	11	15	11	1	37	25	26	14	10	25	40	28	.	32
27	SK Vehicle	2.75	1.07	9	12	16	-8	4	8	1	8	2	-15	-3	-3	7	9	9	24	3	35	27	27	11	8	40	52	36	32	.

N= 401

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 19E: ARMOR CREWMAN

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	Overall Rating	4.62	0.78	.	84	72	58	72	53	69	61	12	20	8	-37	41	12	15	16	4	25	23	22	19	10	27	26	18	22	18	7
2	Eff/Ldr Rating	4.38	0.74	84	.	68	55	76	50	80	65	16	21	8	-32	41	17	26	19	17	31	34	31	27	14	33	32	23	28	23	11
3	Discipline Rtng	4.50	0.83	72	68	.	45	53	41	55	64	-1	12	-14	-35	38	6	15	14	2	21	23	18	17	6	27	18	8	22	14	-2
4	Fitness Rating	4.76	0.82	58	55	45	.	44	39	43	36	10	43	-0	-19	28	8	2	16	-3	1	5	10	5	-4	0	-0	-0	4	2	2
5	Job-Spec Tech	23.19	3.20	72	76	53	44	.	75	71	55	10	14	17	-31	34	23	17	19	13	25	23	26	19	8	27	25	18	15	20	6
6	Job-Spec Other	14.71	1.89	53	50	41	39	75	.	50	41	9	13	13	-18	19	15	9	13	2	6	7	12	2	4	15	12	8	9	9	1
7	Combat Exmplry	8.38	1.36	69	60	55	43	71	50	.	63	15	18	8	-32	34	15	27	15	14	20	23	19	20	7	22	25	19	10	18	2
8	Combat Problems	9.80	1.47	61	65	64	36	55	41	63	.	-1	7	4	-31	29	13	22	13	6	24	18	21	13	8	24	18	17	15	16	-1
9	Awards & Certs	2.52	1.60	12	16	-1	10	10	9	15	-1	.	15	19	-7	13	6	4	-3	13	5	7	-0	10	-2	12	12	3	4	8	8
10	Phys. Readiness	249.41	27.11	20	21	12	43	14	13	18	7	15	.	-1	-10	10	-3	-3	4	2	-6	0	-6	4	1	-4	1	2	-2	2	-7
11	M16 Qualific.	2.40	0.68	8	8	-14	-0	17	13	8	4	19	-1	.	14	-1	7	7	3	10	11	12	13	17	31	10	6	12	2	16	-1
12	Articles 15	0.35	0.77	-37	-32	-35	-19	-31	-18	-32	-31	-7	-10	14	.	-43	-9	-8	-16	1	-13	-17	-17	-7	1	-19	-13	-13	-0	-7	-6
13	Promotion Rate	0.03	0.58	41	41	38	28	34	19	34	29	13	10	-1	-43	.	10	7	15	12	14	24	28	21	2	17	22	18	6	15	1
14	MO Tech.	50.00	9.99	12	17	6	8	23	15	15	13	6	-3	7	-9	10	.	18	24	20	36	27	27	13	18	23	18	9	2	19	0
15	MO Basic	38.16	2.48	15	26	15	2	17	9	27	22	4	-3	7	-8	7	18	.	21	23	30	32	25	21	18	21	25	11	4	19	-0
16	MO Safety	21.85	2.95	16	19	14	16	19	13	15	13	-3	4	3	-16	15	24	21	.	14	22	18	18	10	6	15	13	9	5	17	6
17	MO Comm	28.55	7.59	4	17	2	-3	13	2	14	6	13	2	10	1	12	20	23	14	.	23	28	25	32	11	20	23	13	3	22	3
18	JK Tech.	50.00	9.99	25	31	21	1	25	8	20	24	5	-6	11	-13	14	36	30	22	23	.	60	52	45	34	64	60	44	38	42	7
19	JK Basic	42.16	7.28	23	34	23	5	23	7	23	18	7	0	12	-17	24	27	32	18	28	60	.	65	53	30	65	67	46	41	43	6
20	JK Safety	21.19	4.10	22	31	18	10	26	12	19	21	-0	-6	13	-17	28	27	25	16	25	52	65	.	44	34	46	51	37	26	33	5
21	JK Comm	11.33	3.59	19	27	17	5	19	2	20	13	10	4	17	-7	21	13	21	10	22	45	53	44	.	16	45	51	34	30	24	2
22	JK Identify	10.05	1.78	10	14	6	-4	8	4	7	8	-2	1	31	1	2	18	18	6	11	34	30	34	16	.	24	25	22	18	37	3
23	SK Tech.	54.54	9.66	27	33	27	0	27	15	22	24	12	-4	10	-19	17	23	21	15	20	64	65	46	45	24	.	75	53	59	48	21
24	SK Basic	34.94	8.44	26	32	18	-0	25	12	25	18	12	1	6	-13	22	18	25	13	23	60	67	51	51	28	75	.	68	47	47	12
25	SK Safety	8.18	2.14	18	23	8	-0	18	8	19	17	3	2	12	-13	18	9	11	5	13	44	46	37	34	22	53	58	.	38	33	4
26	SK Comm	7.59	1.80	22	22	22	4	15	9	10	15	4	-2	2	-0	6	2	4	5	3	38	41	26	30	18	59	47	38	.	24	14
27	SK Vehicle	0.54	0.50	7	11	-2	2	6	1	2	-1	8	-7	-1	-6	1	0	-0	6	3	7	6	5	2	3	21	12	4	14	9	.
28	SK Identify	3.01	0.96	18	23	14	2	20	9	18	16	8	2	16	-7	15	19	19	17	23	42	43	33	24	37	48	47	33	24	.	9

N= 335

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 31C: SINGLE CHANNEL RADIO OPERATOR

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	Overall Rating	4.73	0.79	.	83	73	64	74	66	66	66	17	11	2	-31	30	20	24	15	15	-2	24	17	9	14	3	13	19	10	14	2	-2
2	Eff/Ldr Rating	4.48	0.72	83	.	68	57	81	71	68	63	18	12	7	-31	30	24	21	21	15	2	30	28	14	16	6	13	23	12	20	12	4
3	Discipline Rtng	4.64	0.88	73	68	.	52	54	58	53	60	4	4	-11	-32	26	10	14	7	10	-1	20	15	4	15	6	7	9	-4	10	-2	-8
4	Fitness Rating	5.05	0.88	64	57	52	.	47	40	42	42	11	34	-6	-25	24	12	8	10	4	-2	1	2	0	8	-4	6	-5	-6	-2	-9	-12
5	Job-Spec Tech	14.27	2.01	74	81	54	47	.	76	66	57	14	4	5	-16	22	20	24	20	11	-1	29	30	16	15	8	15	19	15	16	13	-1
6	Job-Spec Other	14.37	2.09	66	71	58	40	76	.	54	48	3	-3	-3	-17	22	11	18	15	1	0	17	22	8	9	2	5	5	2	9	3	-9
7	Combat Exmplry	9.09	1.54	66	68	53	42	66	54	.	77	11	1	5	-21	17	6	13	18	23	-7	26	30	15	19	11	12	18	9	14	10	7
8	Combat Problems	10.47	1.71	66	63	60	42	57	48	77	.	9	-1	-2	-22	14	4	16	11	15	-0	22	24	3	14	-1	5	15	5	9	0	-3
9	Awards & Certs	2.16	1.75	17	18	4	11	14	3	11	9	.	23	10	2	12	9	12	6	3	2	10	10	11	-0	-5	8	8	12	4	4	6
10	Phys. Readiness	259.34	29.59	11	12	4	34	4	-3	1	-1	23	.	4	-11	4	1	-10	0	1	-6	-4	-8	4	1	3	-8	-6	-0	1	-13	-5
11	M16 Qualific.	2.16	0.77	2	7	-11	-6	5	-3	5	-2	10	4	.	4	3	4	5	10	7	5	7	10	8	-4	-6	5	9	4	4	11	10
12	Articles 15	0.34	0.84	-31	-31	-32	-25	-16	-17	-21	-22	2	-11	4	.	-34	-9	-3	-7	-12	-3	-16	-9	-13	-20	-10	-3	-11	-4	-12	-4	-3
13	Promotion Rate	-0.02	0.56	30	30	26	24	22	22	17	14	12	4	3	-34	.	8	12	21	9	5	18	17	10	19	13	12	13	15	12	4	-0
14	MO Tech.	78.44	9.49	20	24	10	12	20	11	6	4	9	1	4	-9	8	.	25	25	28	9	42	21	23	21	22	15	39	21	24	9	8
15	MO Basic	21.25	3.84	24	21	14	8	24	18	13	16	12	-10	5	-3	12	25	.	18	27	8	31	31	18	15	5	21	27	24	27	10	15
16	MO Safety	20.15	3.99	15	21	7	10	20	15	18	11	6	0	10	-7	21	25	18	.	23	16	10	21	13	9	6	8	11	10	19	4	9
17	MO Comm	16.73	6.59	15	15	10	4	11	1	23	15	3	1	7	-12	9	28	27	23	.	1	34	29	21	28	21	23	26	17	11	5	25
18	MO Vehicle	11.73	1.31	-2	2	-1	-2	-1	0	-7	-0	2	-6	5	-3	5	9	8	16	1	.	11	9	10	3	-6	7	22	16	14	12	11
19	JK Tech.	57.16	11.68	24	30	20	1	29	17	26	22	10	-4	7	-16	18	42	31	10	34	11	.	60	59	60	37	33	72	49	50	44	25
20	JK Basic	22.12	4.61	17	28	15	2	30	22	30	24	10	-8	10	-9	17	21	31	21	29	9	60	.	58	50	23	21	49	42	43	40	20
21	JK Safety	23.31	4.63	9	14	4	0	16	8	15	3	11	4	8	-13	10	23	18	13	21	10	59	58	.	50	28	30	44	40	48	36	24
22	JK Comm	10.12	2.74	14	16	15	8	15	9	19	14	-0	1	-4	-20	19	21	15	9	38	2	60	50	50	.	32	19	44	36	36	27	19
23	JK Vehicle	4.54	1.82	3	6	6	-4	8	2	11	-1	-5	3	-6	-10	13	22	5	6	21	-6	37	23	28	32	.	17	30	14	16	13	11
24	JK Identify	6.72	2.13	13	13	7	6	15	5	12	5	8	-6	5	-3	12	15	21	8	23	7	33	31	30	19	17	.	27	28	21	11	40
25	SK Tech.	77.87	15.43	19	23	9	-5	19	5	18	15	8	-4	9	-11	13	39	27	11	26	22	72	49	44	44	20	27	.	62	58	48	29
26	SK Basic	10.95	2.74	10	12	-4	-8	15	2	9	5	12	-0	4	-4	15	21	24	10	17	16	49	42	40	36	14	28	62	.	56	42	26
27	SK Safety	11.08	2.81	14	20	10	-2	16	9	14	9	4	1	4	-12	12	24	27	19	11	14	50	43	48	36	16	21	58	56	.	41	27
28	SK Vehicle	3.83	1.84	2	12	-2	-9	13	3	10	0	4	-13	11	-4	4	9	10	4	5	12	44	40	36	27	13	11	48	42	41	.	22
29	SK Identify	1.16	0.93	-2	4	-8	-12	-1	-9	7	-3	6	-5	10	-3	-0	8	15	9	25	11	25	20	24	19	11	40	29	26	27	22	.

N= 239

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 63B: LIGHT WEIGHT VEHICLE MECHANIC

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	Overall Rating	4.55	0.84	.	86	75	57	75	75	68	65	20	7	-4	-24	24	11	-1	5	10	20	15	22	11	21	22	21	15	19
2	Eff/Ldr Rating	4.31	0.83	86	.	75	50	84	78	69	66	21	1	-5	-23	23	19	-1	3	12	23	16	27	18	26	22	19	14	22
3	Discipline Rtng	4.54	0.88	75	75	.	51	63	65	59	66	15	2	-8	-27	26	10	-5	7	9	11	5	19	3	14	23	20	13	14
4	Fitness Rating	4.82	0.86	57	50	51	.	38	49	44	41	13	31	2	-20	20	-2	-2	8	7	-0	2	8	-0	-2	16	14	13	8
5	Job-Spec Tech	22.42	4.10	75	84	63	38	.	78	65	57	21	-1	-5	-16	16	23	1	3	13	28	21	26	19	37	21	18	6	28
6	Job-Spec Other	23.19	3.52	75	78	65	49	78	.	68	55	18	5	-8	-18	17	12	4	4	12	18	17	22	13	21	16	18	9	20
7	Combat Exmplry	8.87	1.61	68	69	59	44	65	68	.	69	14	4	-7	-16	17	13	0	9	9	16	11	23	8	20	18	14	6	13
8	Combat Problems	9.92	1.86	65	66	66	41	57	55	69	.	14	-0	-6	-20	27	10	-3	4	9	17	11	20	7	19	21	18	18	18
9	Awards & Certs	2.31	1.81	20	21	15	13	21	18	14	14	.	4	2	-11	7	11	-5	-0	7	7	2	12	11	13	14	10	8	16
10	Phys. Readiness	255.47	31.93	7	1	2	31	-1	5	4	-0	4	.	10	-10	15	1	8	3	-1	-7	-12	-2	-9	-10	1	0	-3	-4
11	M16 Qualific.	2.19	0.73	-4	-5	-8	2	-5	-8	-7	-6	2	10	.	1	-9	-2	5	-4	-0	-6	3	3	2	-2	-2	2	-0	4
12	Articles 15	0.37	0.85	-24	-23	-27	-20	-16	-18	-16	-20	-11	-10	1	.	-36	-3	-2	-2	-4	-7	-5	-6	-0	-6	-11	-7	-13	-8
13	Promotion Rate	0.04	0.52	24	23	26	20	16	17	17	27	7	15	-9	-36	.	-5	-4	-2	-1	13	9	4	8	13	16	15	9	13
14	MD Tech.	110.11	6.84	11	19	10	-2	23	12	13	10	11	1	-2	-3	-5	.	8	6	18	33	23	19	22	37	19	16	4	24
15	MD Basic	34.96	4.09	-1	-1	-5	-2	1	4	0	-3	-5	8	5	-2	-4	8	.	10	7	6	12	14	12	10	7	15	-1	14
16	MD Safety	21.92	3.25	5	3	7	8	3	4	9	4	-0	3	-4	-2	-2	6	10	.	2	2	5	18	1	1	2	7	-7	-0
17	MD Vehicle	11.22	1.84	10	12	9	7	13	12	9	9	7	-1	-0	-4	-1	18	7	2	.	15	6	4	11	17	6	6	2	13
18	JK Tech.	68.61	11.93	20	23	11	-0	28	18	16	17	7	-7	-6	-7	13	33	6	2	15	.	62	47	62	67	50	39	36	59
19	JK Basic	24.36	4.69	15	16	5	2	21	17	11	11	2	-12	3	-5	9	23	12	5	6	62	.	45	44	47	41	36	22	44
20	JK Safety	18.91	3.05	22	27	19	8	26	22	23	20	12	-2	3	-6	4	19	14	18	4	47	45	.	38	40	36	33	20	39
21	JK Vehicle	15.81	4.03	11	18	3	-0	19	13	8	7	11	-9	2	-0	8	22	12	1	11	62	44	38	.	56	37	31	24	49
22	SK Tech.	56.00	12.89	21	26	14	-2	37	21	20	19	13	-10	-2	-6	13	37	10	1	17	67	47	40	56	.	52	47	30	39
23	SK Basic	16.56	4.24	22	22	23	16	21	18	18	21	14	1	-2	-11	16	19	7	2	6	50	41	36	37	52	.	61	30	56
24	SK Safety	6.02	1.74	21	19	20	14	18	18	14	18	10	0	2	-7	15	16	15	7	6	39	36	33	31	47	61	.	39	50
25	SK Comm	0.90	0.30	15	14	13	13	6	9	8	18	8	-3	-0	-13	8	4	-1	-7	2	36	22	20	24	30	50	39	.	39
26	SK Vehicle	24.10	5.54	19	22	14	8	28	20	13	18	18	-4	4	-6	13	24	14	-0	13	59	44	39	49	69	56	50	39	.

N= 403

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 64C: MOTOR TRANSPORT OPERATOR

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	Overall Rating	4.52	0.78	.	86	78	63	72	59	68	58	13	11	4	-30	33	8	20	15	16	17	19	3	13	12	16	14
2	EFF/Ldr Rating	4.36	0.75	86	.	77	59	78	69	74	58	17	9	6	-25	31	16	20	18	22	22	26	7	21	17	15	21
3	Discipline Rtng	4.53	0.81	78	77	.	52	67	51	58	54	10	3	-2	-29	35	4	14	14	15	15	19	1	16	12	15	16
4	Fitness Rating	4.74	0.87	63	59	52	.	54	39	46	35	3	28	3	-20	21	-2	8	14	5	6	4	2	5	6	7	-2
5	Job-Spec Tech	29.61	3.76	72	78	67	54	.	78	65	52	13	6	7	-21	25	9	16	19	17	20	19	5	16	17	12	15
6	Job-Spec Other	17.79	2.52	59	69	51	39	78	.	63	41	18	4	13	-15	19	12	11	16	17	16	19	4	13	17	7	14
7	Combat Exmplry	8.80	1.45	68	74	58	46	65	63	.	65	12	6	11	-21	22	20	19	16	20	15	20	5	18	8	10	15
8	Combat Problems	9.50	1.63	58	58	54	35	52	41	65	.	8	-3	2	-24	26	12	15	10	16	17	22	7	15	14	20	19
9	Awards & Certs	3.12	2.08	13	17	10	3	13	18	12	8	.	6	11	5	12	8	4	5	-3	-2	1	6	3	4	-1	2
10	Phys. Readiness	248.48	37.70	11	9	3	28	6	4	6	-3	6	.	3	-6	-1	-1	3	2	-4	-4	-4	2	-2	-5	0	-8
11	M16 Qualific.	2.09	0.75	4	6	-2	3	7	13	11	2	11	3	.	4	-5	9	12	5	7	5	3	-0	-1	-3	1	-1
12	Articles 15	0.46	0.98	-30	-25	-29	-20	-21	-15	-21	-24	5	-6	4	.	-36	-1	-11	-11	-7	-13	-12	0	-5	-8	-12	-4
13	Promotion Rate	-0.01	0.57	33	31	35	21	25	19	22	26	12	-1	-5	-36	.	10	9	10	9	12	11	5	11	9	8	11
14	MO Basic	43.44	10.16	8	16	4	-2	9	12	20	12	8	-1	9	-1	10	.	29	10	44	31	30	7	28	21	6	32
15	MO Safety	83.73	9.84	20	20	14	8	16	11	19	15	4	3	13	-11	9	29	.	14	27	31	24	4	24	19	14	24
16	MO Vehicle	33.30	4.19	15	18	14	14	19	16	16	10	5	2	5	-11	10	10	14	.	5	8	15	3	10	11	1	11
17	JK Basic	27.28	5.82	16	23	15	5	17	17	20	16	-3	-4	7	-7	9	44	27	5	.	67	54	10	47	39	20	49
18	JK Safety	33.42	5.42	17	22	15	6	20	16	15	17	-2	-4	5	-13	12	31	31	8	67	.	49	4	42	47	23	49
19	JK Vehicle	35.40	7.70	19	26	19	4	19	19	20	22	1	-4	3	-12	11	30	24	15	54	49	.	11	49	40	27	35
20	JK Identify	2.15	1.41	3	7	1	2	5	4	5	7	6	2	-0	0	5	7	4	3	10	4	11	.	17	10	-2	13
21	SK Basic	16.41	4.36	13	21	16	5	16	13	18	15	3	-2	-1	-5	11	28	24	10	47	42	49	17	.	56	43	68
22	SK Safety	6.44	1.93	12	17	12	6	17	17	8	14	4	-5	-3	-8	9	21	19	11	39	47	40	10	56	.	36	59
23	SK Comm	0.89	0.32	16	15	15	7	12	7	10	20	-1	0	1	-12	8	6	14	1	20	23	27	-2	43	36	.	37
24	SK Vehicle	55.72	10.07	14	21	16	-2	15	14	15	19	2	-8	-1	-4	11	32	24	11	49	49	55	13	68	59	37	.

N= 477

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 71L: ADMINISTRATIVE CLERK

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	Overall Rating	4.92	0.85	.	83	71	57	72	63	63	59	20	24	4	23	20	17	14	3	22	15	17	21	13	11	5	10
2	Eff/Ldr Rating	4.64	0.78	83	.	73	56	73	65	70	60	21	19	2	19	19	25	14	2	29	17	18	28	17	9	7	11
3	Discipline Rtng	5.01	0.88	71	73	.	47	63	55	58	58	13	13	4	27	19	20	10	-3	22	15	11	20	7	8	1	4
4	Fitness Rating	5.23	0.89	57	56	47	.	40	39	35	49	20	35	5	23	20	3	7	-3	1	2	2	-1	0	-5	0	-2
5	Job-Spec Tech	19.88	2.73	72	73	63	40	.	76	54	50	8	7	-5	-21	21	24	8	-2	28	16	16	28	10	9	6	7
6	Job-Spec Other	18.57	3.13	63	65	55	39	76	.	50	46	10	13	-1	-21	17	22	13	1	22	15	16	26	8	9	10	8
7	Combat Exmplry	8.74	1.83	63	70	58	55	54	50	.	72	24	19	8	-15	18	9	20	11	13	23	17	14	13	8	8	23
8	Combat Problems	10.72	1.95	59	60	58	49	50	46	72	.	21	16	7	-22	13	12	14	6	11	26	12	15	13	9	1	14
9	Awards & Certs	2.62	1.73	20	21	13	20	8	10	24	21	.	17	20	-4	9	-0	10	-1	-0	5	11	-0	-2	-2	5	1
10	Phys. Readiness	260.40	33.39	24	19	13	35	7	13	19	16	17	.	11	-9	5	1	6	5	0	-5	8	5	4	12	2	8
11	M16 Qualific.	1.86	0.80	4	2	4	5	-5	-1	8	7	20	11	.	3	2	-4	12	8	-6	7	3	-3	2	-7	-1	13
12	Articles 15	0.22	0.62	-23	-19	-27	-23	-21	-21	-15	-22	-4	-9	3	.	-42	-13	-5	1	-10	-7	2	-10	-5	-5	-5	4
13	Promotion Rate	0.01	0.46	20	19	19	20	21	17	18	13	9	5	2	-42	.	12	5	2	6	6	9	5	7	8	4	-0
14	MO Tech.	86.09	14.26	17	25	20	3	24	22	9	12	-0	1	-4	-13	12	.	28	13	58	34	33	58	25	23	7	11
15	MO Basic	18.56	5.00	14	14	10	7	8	13	20	14	10	6	12	-5	5	28	.	43	29	48	35	23	26	17	6	23
16	MO Safety	20.54	4.00	3	2	-3	-3	-2	1	11	6	-1	5	8	1	2	13	43	.	11	28	23	7	13	10	0	17
17	JK Tech.	42.21	9.53	22	29	22	1	28	22	13	11	-0	0	-6	-10	6	58	29	11	.	47	48	73	42	24	17	17
18	JK Basic	25.23	5.16	15	17	15	2	16	15	23	26	5	-5	7	-7	6	34	46	28	47	.	50	40	44	27	27	28
19	JK Safety	16.24	3.01	17	18	11	2	16	16	17	12	11	8	3	2	9	33	35	23	48	50	.	43	28	32	19	25
20	SK Tech.	44.99	9.78	21	28	20	-1	28	26	14	15	-0	5	-3	-10	5	58	23	7	73	40	43	.	44	33	15	16
21	SK Basic	9.90	2.28	13	17	7	0	10	8	13	13	-2	4	2	-5	7	25	26	13	42	44	38	44	.	32	18	31
22	SK Safety	4.26	1.29	11	9	8	-5	9	9	8	9	-2	12	-7	-5	6	23	17	10	24	27	32	33	32	.	4	15
23	SK Comm	0.38	0.48	5	7	1	0	6	10	8	1	5	2	-1	-5	4	7	6	0	17	27	19	15	18	4	.	11
24	SK Vehicle	2.71	1.21	10	11	4	-2	7	8	23	14	1	8	13	4	-0	11	23	17	17	28	25	16	31	15	11	.

N= 353

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 91A: MEDICAL SPECIALIST

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	Overall Rating	4.61	0.82	.	86	78	60	67	62	71	70	22	15	-2	-29	32	17	6	13	28	24	25	4	8	26	12	15	5
2	Eff/Ldr Rating	4.40	0.77	86	.	76	56	73	67	73	71	24	13	-4	-30	33	20	9	19	26	25	21	-2	13	33	14	16	9
3	Discipline Rtng	4.54	0.91	78	76	.	47	60	47	58	69	12	7	-8	-29	31	15	11	13	28	21	20	-4	6	33	14	11	10
4	Fitness Rating	4.74	0.92	60	56	47	.	41	38	49	47	10	39	0	-20	18	3	-0	-0	3	7	4	7	1	-1	2	-4	-15
5	Job-Spec Tech	23.09	3.24	67	73	60	41	.	67	55	54	15	6	-1	-27	26	18	2	13	22	16	14	-3	3	32	5	15	7
6	Job-Spec Other	18.47	2.55	62	67	47	38	67	.	64	51	28	7	9	-17	27	10	6	16	18	25	20	5	15	23	11	16	16
7	Combat Exmplry	9.20	1.48	71	73	58	49	55	64	.	79	30	9	9	-20	26	16	10	15	22	25	22	1	18	28	20	17	12
8	Combat Problems	10.11	1.77	70	71	69	47	54	51	79	.	23	5	-5	-28	30	14	6	11	24	22	23	-1	9	32	25	16	12
9	Awards & Certs	3.04	2.01	22	24	12	10	15	28	30	23	.	14	34	-6	13	3	7	22	4	10	8	11	18	4	11	12	6
10	Phys. Readiness	255.71	31.94	15	13	7	39	6	7	9	5	14	.	17	-11	-2	4	-6	-5	-3	-7	-2	3	-5	-6	-3	-6	-7
11	M16 Qualific.	2.08	0.78	-2	-4	-8	0	-1	9	9	-5	34	17	.	-1	-4	3	0	8	-8	5	-7	-0	12	-4	-2	2	2
12	Articles 15	0.41	0.89	-29	-30	-29	-20	-27	-17	-20	-28	-8	-11	-1	.	-33	-10	1	-7	-10	-7	-6	12	-5	-13	-15	-6	-1
13	Promotion Rate	-0.00	0.58	32	33	31	18	26	27	26	30	13	-2	-4	-33	.	10	9	7	16	20	9	-9	11	18	11	14	11
14	MD Tech.	50.48	10.02	17	20	15	3	18	10	16	14	3	4	3	-10	10	.	16	34	39	27	30	2	13	44	8	28	14
15	MD Basic	9.57	3.00	6	9	11	-0	2	6	10	6	7	-6	0	1	9	16	.	17	21	37	21	9	14	17	18	22	11
16	MD Safety	35.52	4.30	13	19	13	-0	13	16	15	11	22	-5	2	-7	7	34	17	.	32	32	33	3	17	30	10	33	13
17	JK Tech.	85.32	13.71	28	26	28	3	22	18	22	24	4	-3	-8	-10	16	39	21	32	.	54	78	12	16	47	20	48	22
18	JK Basic	15.19	3.63	24	25	21	7	16	25	25	22	10	-7	5	-7	20	27	37	32	54	.	55	8	24	41	23	38	22
19	JK Safety	42.71	7.35	25	21	20	4	14	20	22	23	8	-2	-7	-6	9	30	21	33	78	55	.	12	16	55	21	49	21
20	JK Vehicle	2.42	1.04	4	-2	-4	7	-3	5	1	-1	11	3	-0	12	-9	2	9	3	13	8	12	.	10	2	-3	6	6
21	JK Identify	6.62	2.32	8	13	6	1	3	15	18	9	18	-5	12	-5	11	13	14	17	16	24	16	10	.	15	15	13	13
22	SK Tech.	91.65	17.57	28	33	33	-1	32	23	29	32	4	-8	-4	-13	18	44	17	30	67	41	55	2	15	.	24	52	36
23	SK Basic	2.04	0.78	12	14	14	2	5	11	20	28	11	-3	-2	-16	11	8	18	10	20	23	21	-3	15	24	.	26	14
24	SK Safety	5.77	1.56	15	16	11	-4	15	18	17	16	12	-8	2	-6	14	28	22	33	48	38	49	6	13	52	26	.	27
25	SK Vehicle	4.51	1.62	6	9	10	-15	7	16	12	12	6	-7	2	-1	11	14	11	18	22	22	21	6	13	36	14	27	.

N= 372

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 95B: MILITARY POLICE

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	Overall Rating	4.74	0.80	.	87	69	70	78	68	74	70	18	22	13	-28	21	15	18	8	4	1	12	10	8	4	9	8	19	8	7	6	8	-6
2	Eff/Ldr Rating	4.50	0.73	87	.	71	61	77	72	76	68	20	17	11	-22	19	14	21	10	10	1	10	13	7	7	15	10	18	13	12	8	9	-5
3	Discipline Rtng	4.71	0.77	69	71	.	46	65	48	55	63	6	7	4	-27	26	6	9	5	7	-3	11	10	6	8	12	12	15	18	14	6	10	-3
4	Fitness Rating	4.90	0.84	70	61	46	.	58	56	55	52	16	43	13	-26	16	9	12	7	5	2	0	3	-2	-0	3	-3	3	5	2	-1	1	-7
5	Job-Spec Tech	29.00	3.66	78	77	65	58	.	73	68	63	15	15	11	-19	17	12	19	8	12	2	14	16	9	5	9	7	17	11	12	5	5	-2
6	Job-Spec Other	23.60	3.10	68	72	48	56	73	.	71	61	32	18	27	-16	8	11	22	14	19	7	2	13	6	4	16	7	9	12	6	5	4	-2
7	Combat Exmplry	9.56	1.36	74	78	55	55	68	71	.	79	19	19	16	-17	14	17	19	14	14	6	10	17	11	7	15	6	19	15	9	9	8	1
8	Combat Problems	10.45	1.53	70	68	63	52	63	61	79	.	15	15	15	-28	21	14	16	11	10	-0	16	18	17	11	14	10	19	15	9	6	13	0
9	Awards & Certs	3.17	2.09	18	20	6	16	15	32	19	15	.	20	26	-3	11	8	16	6	8	11	-11	2	-1	7	-0	9	4	4	10	4	9	-0
10	Phys. Readiness	251.75	32.78	22	17	7	43	15	18	19	15	20	.	13	-12	7	-1	6	4	2	4	-6	1	-3	-3	-2	-12	-2	-8	-4	-3	2	-5
11	M16 Qualific.	2.28	0.76	13	11	4	13	11	27	16	15	26	13	.	1	-3	4	6	5	4	6	-3	7	3	2	-9	-1	1	-0	-2	-2	-1	4
12	Articles 15	0.27	0.70	-28	-22	-27	-26	-19	-16	-17	-28	-3	-12	1	.	-39	-4	-0	-8	-3	2	-8	-8	-5	-2	0	0	-7	-6	-3	5	-7	6
13	Promotion Rate	0.01	0.47	21	19	26	16	17	8	14	21	11	7	-3	-39	.	4	4	6	1	-3	15	15	16	10	6	2	0	10	7	1	2	-1
14	MO Tech.	31.58	4.63	15	14	6	9	12	11	17	14	8	-1	4	-4	4	.	18	12	6	11	13	11	10	7	3	5	14	7	3	10	6	-0
15	MO Basic	50.04	10.28	18	21	9	12	19	22	19	16	16	6	6	-0	4	18	.	20	21	18	18	34	26	21	17	5	12	27	23	12	15	-7
16	MO Safety	31.76	5.16	8	10	5	7	8	14	14	11	6	4	5	-8	6	12	20	.	9	15	10	20	21	21	12	9	15	17	19	9	10	-6
17	MO Comm	10.57	2.17	4	10	7	5	12	19	14	10	8	2	4	-3	1	6	21	9	.	31	14	21	13	30	14	7	9	21	16	17	11	-10
18	MO Vehicle	10.56	1.63	1	1	-3	2	2	7	6	-0	11	4	6	2	-3	11	18	15	31	.	1	4	8	19	16	2	11	12	9	13	10	-1
19	JK Tech.	38.44	5.90	12	10	11	0	14	2	10	16	-11	-6	-3	-8	15	13	18	10	14	1	.	60	53	35	18	15	40	33	28	24	19	1
20	JK Basic	50.11	9.99	10	13	10	3	16	13	17	18	2	1	7	-8	15	11	34	20	21	4	60	.	60	51	32	22	38	49	46	35	31	4
21	JK Safety	25.52	4.55	8	7	6	-2	9	6	11	17	-1	-3	3	-5	16	10	26	21	13	8	53	60	.	40	24	20	36	37	38	27	28	0
22	JK Comm	13.54	4.62	4	7	8	-0	5	4	7	11	7	-3	2	-2	10	7	21	21	30	19	35	51	40	.	26	18	22	33	31	36	24	-3
23	JK Vehicle	2.03	1.19	9	15	12	3	9	16	15	14	-0	-2	-9	0	6	3	17	12	14	16	18	32	24	26	.	15	18	23	23	20	17	-4
24	JK Identify	6.88	2.29	8	10	12	-3	7	7	6	10	9	-12	-1	0	2	5	5	9	7	2	15	22	20	18	15	.	21	20	21	17	12	-2
25	SK Tech.	40.20	7.04	19	18	15	3	17	9	19	19	4	-2	1	-7	0	14	12	15	9	11	40	38	36	22	18	21	.	49	49	38	37	2
26	SK Basic	17.85	3.86	8	13	18	5	11	12	15	15	4	-8	-0	-6	10	7	27	17	21	12	33	49	37	33	23	20	49	.	60	40	44	-1
27	SK Safety	14.45	3.35	7	12	14	2	12	6	9	9	10	-4	-2	-3	7	3	23	18	16	9	28	46	38	31	23	21	49	60	.	39	40	-6
28	SK Comm	3.12	1.23	6	8	6	-1	5	5	9	6	4	-3	-2	5	1	10	12	9	17	13	24	35	27	36	20	17	38	40	39	.	32	-0
29	SK Vehicle	6.02	1.90	8	9	10	1	5	4	8	13	9	2	-1	-7	2	6	15	10	11	10	19	31	28	24	17	12	37	44	40	32	.	-1
30	SK Identify	0.29	0.51	-6	-5	-3	-7	-2	-2	1	0	-0	-5	4	6	-1	-0	-7	-6	-10	-1	1	4	0	-3	-4	-2	2	-1	-6	-0	-1	.

N= 506

SUMMARY OF RESULTS

FROM THE WITHIN-MOS PRINCIPAL COMPONENT ANALYSES

OF THE PERFORMANCE FACTOR SCORES

- Emergence of method factors
  - -- Written knowledge tests
  - -- Rating scales
- Correspondence between Army-wide BARS and administrative measures factors
  - -- Effort and Leadership from the Army-wide BARS and Letters and Certificates from the administrative measures
  - -- Personal Discipline from the Army-wide BARS and Articles 15/Flag Actions and Promotion Rate Deviation Score from the administrative measures
  - -- Physical Fitness and Military Bearing from the Army-wide BARS and Physical Readiness Test Score from the administrative measures
- Lack of correspondence between performance test scores (i.e., measures of "maximal" performance) and performance ratings (i.e., measures of "typical" performance)
- Distinction between technical (MOS-specific) factor and remaining general soldiering factors
- Lack of distinction among different general soldiering factors (i.e., Communications, Vehicle Operation and Maintenance, etc.)

LATENT CONSTRUCTS

UNDERLYING THE PROJECT A ENLISTED PERFORMANCE MEASURES

- "Content" constructs
  - -- Core Technical Proficiency
  - -- General Soldiering Proficiency
  - -- Effort and Leadership
  - -- Personal Discipline
  - -- Physical Fitness and Military Bearing
- "Method" constructs
  - -- Written knowledge tests
  - -- Rating scales

DEFINITIONS OF THE PERFORMANCE CONSTRUCTS

- Core Technical Proficiency
  This performance construct represents the proficiency with which the soldier performs the tasks that are "central" to the MOS. The tasks represent the core of the job and they are the primary definers of the MOS. For example, the first tour Armor Crewman starts and stops the tank engines; prepares the loader's station; loads and unloads the main gun; boresights the M60A3; engages targets with the main gun; and performs misfire procedures. This performance construct does not include the individual's willingness to perform the task or the degree to which the individual can coordinate efforts with others. It refers to how well the individual can execute the core technical tasks the job requires, given a willingness to do so.

- General Soldiering Proficiency
  In addition to the core technical content specific to an MOS, individuals in every MOS also are responsible for being able to perform a variety of general soldiering tasks -- for example, determines grid coordinates on military maps; puts on, wears and removes M17 series protective mask with hood; determines a magnetic azimuth using a compass; collects/reports information - SALUTE; and recognizes and identifies friendly and threat aircraft. Performance on this construct represents overall proficiency on these general soldiering tasks. Again, it refers to how well the individual can execute general soldering tasks, given a willingness to do so.

- Effort and Leadership
  This performance construct 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. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient? While appropriate knowledges and skills are necessary for successful performance, this construct is only meant to reflect the individual's willingness to do the job required and to be cooperative and supportive with other soldiers.

- Personal Discipline
  This performance construct reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems. People who rank high on this construct show a commitment to high standards of personal conduct.

- Physical Fitness and Military Bearing
  This performance construct represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

MAPPING OF PERFORMANCE FACTORS ONTO LATENT PERFORMANCE CONSTRUCTS

	Latent Performance Constructs							
	Content Constructs					Method Constructs		M16
Performance Factor	Core Proficiency	General Soldier Proficiency	Effort and Leadership	Personal Discipline	Physical Fitness/ Military Bearing	Written Knowledge Tests	Rating Scales	M16 Qualification
AMB Effort			X				X	
AMB Discipline				X			X	
AMB Fitness					X		X	
AMB Overall			X	X			X	
MOS Core	X		X				X	
MOS Other			X				X	
Cmbt Perform Well			X				X	
Cmbt Avoid Mistake			X	X			X	
Adm Letters/Certs			X					
Adm Phys Readiness					X			
Adm M16								X
Adm Articles 15				X				
Adm Promotion Rate				X				
MO Technical	X							
MO Communication		X						
MO Vehicles		X						
MO General Soldier		X						
MO ID Threat/Target		X						
MO Safety/Survival		X						
JK Technical	X					X		
JK Communications		X				X		
JK Vehicles		X				X		
JK General Soldier		X				X		
JK ID Threat/Target		X				X		
JK Safety/Survival		X				X		
SK Technical	X					X		
SK Communications		X				X		
SK Vehicles		X				X		
SK General Soldier		X				X		
SK ID Threat/Target		X				X		
SK Safety/Survival		X				X		

Note: Within each rating instrument, all of the factors were constrained to have an equal loading on the Rating Scales method construct. For example, the Perform Well and Avoid Mistakes factors from the Combat Performance Prediction Scales were constrained to have identical loadings on the Rating Scales method construct, but this loading did not have to be the same as the loading for the Army-Wide BARS factors, the MOS-Specific

UNIQUENESS ESTIMATES

SEPARATE MODEL FOR EACH JOB

	Military Occupational Specialty								
Factor Score	11B	13B	19E	31C	63B	64C	71L	91A	95B
HO Tech	--	.52	.71	.48	.64	.74	.33	.57	.88
HO Soldier	.59	.66	.75	.52	.95	.74	.55	.76	.63
HO Safety	.92	.85	.75	.52	.95	.59	.79	.71	.77
HO Comm	.95	.95	.81	.62	--	--	--	--	.82
HO Vehicle	--	--	--	.03	.95	**	--	--	.90
JK Tech	--	.21	.30	.15	.12	.39	.17	.11	.53
JK Soldier	.10	.43	.22	.26	.29	.74	.31	.58	.43
JK Safety	.32	.53	.32	.31	.45	.49	.44	.15	.57
JK Comm	.56	.93	.32	.34	--	--	--	--	.64
JK Vehicle	--	--	--	.56	.32	**	--	.94	.82
JK Identify	.36	.89	.40	.51	--	.95	--	.92	.90
SK Tech	--	.27	.13	.09	.10	.14	.14	.15	.52
SK Soldier	.09	.37	.14	.48	.31	.42	.54	.74	.46
SK Safety	.46	.59	.43	.41	.50	.55	.72	.47	.55
SK Comm	.40	.72	.35	--	.65	.82	.78	--	.67
SK Vehicle	.73	.62	.69	.55	.18	**	.73	.76	.75
SK Identify	--	.45	.10	.22	.25	.25	.34	.10	.13
Overall Rating	.13	.13	.13	.13	.13	.13	.13	.13	.18
Eff/Ldr Rating	.11	.11	.11	.11	.11	.05	.11	.11	.05
Discpln Rating	.22	.22	.22	.22	.22	.05	.22	.22	.06
Fitness Rating	.38	.38	.38	.38	.38	.05	.38	.38	.05
MOS Tech Rtngs	.08	.11	.13	.14	.08	.37	.17	.12	.33
MOS Other Rtng	.10	.13	.17	.19	.12	.35	.20	.18	.27
Comb Exmplry	.02	.02	.02	.02	.02	.14	.02	.02	.08
Comb Problems	.13	.13	.13	.13	.13	.60	.13	.13	.40
Awards/Cert	.89	.94	.93	.95	.91	.94	.86	.85	.90
Phys Readiness	.95	.33	.67	.34	.50	.83	.46	.49	.49
Articles 15	.58	.59	.68	.60	.56	.76	.51	.75	.64
Promotion Rate	.45	.60	.53	.41	.57	.64	.62	.67	.70
M16	.50	.50	.50	.50	.50	.50	.50	.50	.50

** Vehicle content was merged into the Technical factor for 64C.

GOODNESS-OF-FIT INDICES

SEPARATE MODEL FOR EACH JOB

	MOS	Root Mean Square Residual	Chi-Square	df	p
11B:	Infantryman	.061	326.2	227	.02
13B:	Cannon Crewman	.057	350.0	322	.14
19E:	Tank Crewman	.065	170.0	348	.999
31C:	Radio/Teletype Operator	.069	369.2	375	.58
63B:	Vehicle/Generator Mechanic	.060	332.1	296	.07
64C:	Motor Transport Operator	.058	280.1	247	.07
71L:	Administrative Clerk	.067	232.6	249	.77
91A:	Medical Specialist	.061	277.1	275	.45
95B:	Military Police	.052	470.0	374	.001

FACTOR LOADINGS

SEPARATE MODEL FOR EACH JOB

	Military Occupational Specialty								
Construct/Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Core Technical									
HO Tech	--	.61	.47	.64	.51	.29	.77	.59	.32
JK Tech	--	.75	.78	.79	.74	.26	.78	.75	.32
SK Tech	--	.70	.79	.73	.82	.55	.229	.81	.43
MOS Tech Rtng	--	.45	.10	.22	.25	.25	.34	.10	.13
General Soldiering									
HO Soldier	.60	.51	.46	.64	.17	.50	.60	.42	.60
HO Safety	.26	.33	.32	.31	.12	.63	.37	.48	.47
HO Comm	.05	.06	.39	.56	--	--	--	--	.80
HO Vehicle	--	--	--	.22	.17	**	--	--	.31
JK Soldier	.76	.52	.74	.62	.45	.48	.87	.58	.46
JK Safety	.55	.37	.75	.38	.71	.51	.72	.58	.33
JK Comm	.30	.23	.66	.38	--	--	--	--	.29
JK Vehicle	--	.17	--	.10	.41	**	--	--	.35
JK Identify	.46	--	.20	.28	--	.12	--	.24	.21
SK Soldier	.73	.45	.67	.39	.78	.56	.45	.44	.42
SK Safety	.47	.32	.53	.62	.57	.47	.30	.64	.32
SK Comm	.42	.26	.42	--	.41	.35	.20	--	.20
SK Vehicle	.22	.24	.05	.30	.61	**	.22	.47	.28
SK Identify	.46	--	.46	.13	--	--	--	--	--
Effort/Leadership									
Eff/Ldr Rating	.76	.56	.85	.64	.68	.83	.66	.76	.70
MOS Tech Rtngs	.70	--	.63	.40	.41	.50	.25	.59	.52
MOS Other Rtng	.77	.41	.48	.43	.54	.62	.43	.61	.56
Comb Emplry	.80	.47	.68	.54	.57	.87	.63	.80	.77
Comb Problems	.48	.20	--	.39	.52	.53	.55	--	.56
Awards/Cert	.32	.23	.24	.19	.28	.25	.34	.34	.22
Overall Rating	.46	.39	.33	.17	.57	.42	.65	--	.41

FACTOR LOADINGS

SEPARATE MODEL FOR EACH JOB

(continued)

	Military Occupational Specialty								
Construct/Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Discipline									
Discpln Rtng	.77	.58	.73	.45	.63	.85	.74	.58	.73
Comb Problems	.29	.16	.62	.03	.05	.19	--	.82	.33
Articles 15	-.63	-.61	-.55	-.62	-.65	-.47	-.69	-.46	-.60
Promotion Rate	.74	.61	.68	.79	.63	.57	.59	.54	.54
Overall Rating	.39	.20	.53	.54	.09	.42	.06	.75	.38
Fitness/Bearing									
Fitness Ratngs	.69	.23	.84	.48	.54	.42	.50	.60	.78
Phys Readiness	.11	.90	.49	.89	.70	.53	.76	.69	.69
Ratings Method									
AW Ratings	.60	.73	.47	.70	.66	.54	.65	.66	.66
MOS Ratings	.73	.73	.60	.69	.67	.49	.69	.54	.63
Comb Ratings	.47	.65	.55	.69	.57	.27	.55	.47	.40
Written Method									
JK Tech	--	.47	.28	.55	.59	.73	.44	.58	.57
JK Soldier	.41	.51	.33	.40	.61	.57	.11	.37	.59
JK Safety	.37	.52	.12	.63	.08	.49	.17	.76	.57
JK Comm	.34	.11	.07	.55	--	--	--	--	.52
JK Vehicle	--	--	--	.42	.62	**	--	.24	.21
JK Identify	-.15	.23	.50	.36	--	.05	--	.08	.23
SK Tech	--	.48	.48	.55	.46	.88	.42	.27	.50
SK Soldier	.50	.66	.54	.59	.15	.51	.54	--	.54
SK Safety	.53	.55	.42	.29	.34	.48	.44	.19	.60
SK Comm	.51	.47	.46	--	.16	.24	.05	--	.42
SK Vehicle	.49	.57	.24	.48	.55	**	.38	.05	.42
SK Identify	.21	--	.42	.44	--	--	--	--	--
M16 Qualification	.71	.71	.71	.71	.71	.71	.71	.71	.71

** Vehicle content was merged into the Core Technical factor for 64C.

ESTIMATED CONSTRUCT CORRELATIONS

SEPARATE MODEL FOR EACH JOB

		Military Occupational Specialty							
1st Construct	2nd Construct	11B	13B	19E	31C	63B	64C	71L	91A	95B
Core Technical	Gen Soldiering	--	.77	.83	.63	.58	.73	.48	.66	.70
	Effort/Lead	.67	.86	.51	.44	.50	.78	.44	.35	.46
	Discipline	.42	.13	.37	.26	.12	.69	.19	.43	.50
	Fitness	.25	.01	.03	.04	-.18	-.09	.10	-.05	-.09
	M16	.27	.00	.04	.11	.05	.05	-.09	-.17	-.10
General Soldiering	Effort/Lead	--	.89	.58	.57	.53	.44	.37	.43	.40
	Discipline	--	.29	.45	.30	.29	.29	.04	.37	.24
	Fitness	--	-.19	.05	-.05	-.03	-.14	.09	-.05	.00
	M16	--	-.06	.30	.30	.04	.11	.27	.02	.02
Effort/ Leadership	Discipline	.49	.67	.62	.55	.65	.51	.51	.59	.39
	Fitness	.57	.04	.38	-.11	.10	.23	.32	.21	.42
	M16	.38	-.13	.21	.24	-.02	.35	.22	.17	.28
Discipline	Fitness	.33	.05	.24	.24	.30	.30	.27	.19	.25
	M16	-.12	-.25	-.30	.09	-.28	-.11	.01	-.28	-.08
Fitness	M16	.52	.26	-.05	.02	.19	.22	.18	.27	.26

TESTING THE LATENT STRUCTURE MODEL

ACROSS ALL NINE MOS SIMULTANEOUSLY:

ASSUMPTIONS

- Intercorrelations among the performance constructs are the same for all MOS

- The loadings of the Army-wide factors (i.e., the Army-wide BARS factors, the combat factors, and the administrative measures "factors") on the content and method constructs are constant across MOS

- No M16 factor or construct

UNIQUENESS ESTIMATES

SINGLE MODEL ACROSS ALL JOBS

	Military Occupational Specialty								
Factor Score	11B	13B	19E	31C	63B	64C	71L	91A	95B
HO Tech	--	.62	.79	.62	.76	.91	.44	.68	.90
HO Soldier	.72	.58	.80	.70	.95	.73	.64	.87	.67
HO Safety	.95	.84	.90	.87	.95	.73	.90	.75	.81
HO Comm	.95	.95	.86	.71	--	--	--	--	.82
HO Vehicle	--	--	--	.95	.95	**	--	--	.93
JK Tech	--	.23	.28	.13	.15	.32	.28	.16	.60
JK Soldier	.10	.44	.28	.40	.48	.41	.44	.47	.40
JK Safety	.48	.56	.41	.49	.62	.44	.55	.26	.54
JK Comm	.85	.91	.57	.55	--	--	--	--	.67
JK Vehicle	--	--	--	.87	.44	**	--	.95	.85
JK Identify	.71	.90	.84	.81	--	.95	--	.64	.90
SK Tech	--	.25	.10	.24	.18	.17	.27	.19	.54
SK Soldier	.13	.37	.20	.52	.41	.31	.58	.83	.49
SK Safety	.54	.62	.54	.51	.55	.51	.80	.29	.54
SK Comm	.46	.75	.48	--	.77	.78	.92	--	.70
SK Vehicle	.75	.68	.95	.61	.31	**	.86	.86	.75
Overall Rating*	.18	.18	.18	.18	.18	.18	.18	.18	.18
Eff/Ldr Rating*	.09	.09	.09	.09	.09	.09	.09	.09	.09
Discpln Rating*	.17	.17	.17	.17	.17	.17	.17	.17	.17
Fitness Rating*	.05	.05	.05	.05	.05	.05	.05	.05	.05
MOS Tech Rtngs*	.18	.34	.22	.24	.18	.18	.18	.18	.25
MOS Other Rtng*	.05	.24	.46	.37	.05	.05	.05	.05	.27
Comb Exmplry*	.26	.26	.26	.26	.26	.26	.26	.26	.26
Comb Problems*	.29	.29	.29	.29	.29	.29	.29	.29	.29
Awards/Cert*	.93	.93	.93	.93	.93	.93	.93	.93	.93
Phys Readiness*	.83	.83	.83	.83	.83	.83	.83	.83	.83
Articles 15*	.77	.77	.77	.77	.77	.77	.77	.77	.77
Promotion Rate*	.70	.70	.70	.70	.70	.70	.70	.70	.70

* These loadings were constrained to be equal across all MOS.
** Vehicle content was merged into the Core Technical factor for 64C.

GOODNESS OF FIT INDICES

SINGLE MODEL ACROSS ALL JOBS

- Chi-square = 2508.1
  df = 2403
  p = .07

- Root Mean Square Residual = .047

FACTOR LOADINGS

SINGLE MODEL ACROSS ALL JOBS

	Military Occupational Specialty								
Construct/Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Core Technical									
HO Tech	--	.59	.43	.58	.46	.27	.71	.54	.29
JK Tech	--	.71	.79	.76	.57	.72	.70	.74	.37
SK Tech	--	.66	.70	.54	.73	.55	.68	.85	.42
MOS Tech Rtng	--	.21	.12	.16	.25	.01	.12	.05	-.02
General Soldiering									
HO Soldier	.52	.66	.44	.52	.16	.51	.57	.35	.58
HO Safety	.20	.44	.31	.36	.10	.49	.30	.50	.41
HO Comm	.06	.12	.37	.52	--	--	--	--	.43
HO Vehicle	--	--	--	.15	.21	**	--	--	.27
JK Soldier	.95	.50	.79	.64	.42	.69	.66	.69	.49
JK Safety	.69	.36	.75	.45	.53	.66	.57	.65	.42
JK Comm	.35	.25	.59	.51	--	--	--	--	.39
JK Vehicle	--	--	--	.28	.37	**	--	.07	.34
JK Identify	.43	.21	.34	.36	--	.12	--	.39	.18
SK Soldier	.81	.40	.67	.33	.70	.50	.42	.40	.38
SK Safety	.57	.34	.45	.40	.63	.43	.31	.62	.34
SK Comm	.51	.21	.31	--	.42	.29	.17	--	.23
SK Vehicle	.35	.22	.06	.17	.65	**	.32	.36	.21
Effort/Leadership									
Eff/Ldr Rating*	.76	.76	.76	.76	.76	.76	.76	.76	.76
MOS Tech Rtngs*	.59	.33	.54	.50	.45	.62	.43	.62	.62
MOS Other Rtng*	.77	.59	.33	.45	.59	.48	.47	.58	.58
Comb Emplry*	.72	.72	.72	.72	.72	.72	.72	.72	.72
Comb Problem*	.44	.44	.44	.44	.44	.44	.44	.44	.44
Awards/Cert*	.26	.26	.26	.26	.26	.26	.26	.26	.26
Overall Rating*	.48	.48	.48	.48	.48	.48	.48	.48	.48

FACTOR LOADINGS

SINGLE MODEL ACROSS ALL JOBS

(continued)

	Military Occupational Specialty								
Construct/Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Discipline									
Discpln Rtng*	.69	.69	.69	.69	.69	.69	.69	.69	.69
Comb Problems*	.25	.25	.25	.25	.25	.25	.25	.25	.25
Articles 15*	-.48	-.48	-.48	-.48	-.48	-.48	-.48	-.48	-.48
Promotion Rate*	.52	.52	.52	.52	.52	.52	.52	.52	.52
Overall Rating*	.28	.28	.28	.28	.28	.28	.28	.28	.28
Fitness/Bearing									
Fitness Ratngs*	.82	.82	.82	.82	.82	.82	.82	.82	.82
Phys Readiness*	.37	.37	.37	.37	.37	.37	.37	.37	.37
Ratings Method									
AW Ratings*	.56	.56	.56	.56	.56	.56	.56	.56	.56
MOS Ratings*	.61	.61	.61	.61	.61	.61	.61	.61	.61
Comb Ratings*	.42	.42	.42	.42	.42	.42	.42	.42	.42
Written Method									
JK Tech	--	.49	.29	.54	.71	.30	.42	.49	.49
JK Soldier	-.16	.51	.29	.40	.53	.25	.28	.60	.60
JK Safety	-.07	.49	.07	.52	.26	.28	.35	.52	.52
JK Comm	.00	.11	.19	.38	--	--	--	.41	.41
JK Vehicle	--	--	--	.19	.62	**	--	.20	.20
JK Identify	-.05	.20	.12	.17	--	.10	--	.25	.25
SK Tech	--	.54	.65	.64	.49	.71	.45	.53	.53
SK Soldier	.44	.68	.58	.61	.25	.66	.50	.60	.60
SK Safety	.34	.51	.49	.57	.18	.56	.30	.59	.59
SK Comm	.51	.46	.60	--	.20	.36	.20	.50	.50
SK Vehicle	.38	.51	.17	.60	.45	**	.17	.46	.46

* These loadings were constrained to be equal across all MOS.
** Vehicle content was merged into the Core Technical factor for 64C.

ESTIMATED PERFORMANCE CONSTRUCT CORRELATIONS

SINGLE MODEL ACROSS ALL JOBS

First Construct	Second Construct	Correlation
Core Technical	General Soldiering	.80
	Effort/Leadership	.48
	Discipline	.35
	Fitness/Bearing	.01
General Soldiering	Effort Leadership	.47
	Discipline	.35
	Fitness/Bearing	.06
Effort/Leadership	Discipline	.67
	Fitness/Bearing	.42
Discipline	Fitness/Bearing	.40

SCORING THE CONSTRUCTS:

PRELIMINARY DECISIONS

- "Rational" vs. regression weights

- Regression weights were not used because:

  -- They would be difficult to explain

  -- We were concerned about their stability

- Mapping of factors onto constructs

  Each factor was assigned to one construct. If a factor was assigned to two constructs in the latent structure model, for scoring purposes it was assigned to the construct on which it had the highest loading.

- Unit weights by measurement method

  As an intermediate step in computing construct scores, we first computed construct subscores by combining all of the factors from a given measurement method. We then summed these subscores to compute the total construct score.

CONSTRUCT SUBSCORES

- Core Technical Proficiency
  - -- Hands-On Subscore: Average percent GO across all Core Technical tasks
  - -- Knowledge Subscore: Sum of job and school knowledge Core Technical items answered correctly

- General Soldiering Proficiency
  - -- Hands-On Subscore: Average percent GO across all General Soldiering tasks
  - -- Knowledge Subscore: Sum of job knowledge and school knowledge General Soldiering items answered correctly

- Effort and Leadership
  - -- Overall Effectiveness Subscore: Overall Effectiveness rating from the Army-wide BARS
  - -- BARS Subscore: Sum of (1) the Effort and Leadership factor from the Army-wide BARS, and (2) the Core factor and (3) the Other factor from the MOS-specific BARS
  - -- Combat Subscore: Average rating across all of the items from the Combat Performance Prediction Scales
  - -- Administrative Measures Subscore: Letters and Certificates factor score from the administrative measures

- Personal Discipline
  - -- Army-wide BARS Subscore: Personal Discipline factor score from the Army-wide BARS
  - -- Administrative Measures Subscore: Sum of (1) the Articles 15/Flag Actions and (2) the Promotion Rate Deviation factor scores from the administrative measures

- Physical Fitness and Military Bearing
  - -- Army-wide BARS Subscore: Physical Fitness and Military Bearing factor score from the Army-wide BARS
  - -- Administrative Measures Subscore: Physical Readiness Test factor score from the administrative measures

PROJECT A CONCURRENT VALIDATION:

TREATMENT OF MISSING DATA

Lauress L. Wise
Jeffrey J. McHenry
Winnie Y. Young

American Institutes for Research

Presented at a Data Analysis Workshop of the
Committee on Performance of Military Personnel

Baltimore

December 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

PROJECT A CONCURRENT VALIDATION: TREATMENT OF MISSING DATA

The job performance data collected in the Project A Concurrent Validation (CV) are unprecedented in scope. The data cover 19 diverse Military Occupational Specialties (MOS) and were collected using an exhaustive array of different job performance measurement methods. For nine MOS, designated Batch A, a complete set of performance measures was developed and administered. For the remaining MOS, designated Batch Z, an abbreviated set of "Army-wide" measures was used.

The measures that we used included:

- hands-on tests (HO): observation and scoring of performance on 15 carefully sampled job tasks
- written tests of job knowledge (JK): tests of facts and procedures for 30 carefully sampled job tasks
- written tests of school knowledge (SK): tests of facts and procedures taught during training for the MOS
- ratings of performance by peers and supervisors on several sets of rating scales, including:
  - 11 Army-wide Behavior Summary Scales (AWB)
  - 7 to 13 MOS-specific Behavior Summary Scales (MSB)
  - 15 Job Task Rating Scales (JTR) for Batch A MOS
  - 11 Common Task Rating Scales (CTR) for Batch Z MOS
  - 40 Combat Performance Prediction Scales (CPP)
- self-report of administrative and personnel records (ADM), including:
  - letters and commendations
  - Physical Readiness Test Score
  - Marksmanship Score
  - disciplinary actions (Articles 15, Flag Actions)
- data from operational Army files on promotion rates and scores on Skill Qualification Tests.

Data on possible moderators of job performance (job history and work environment) and on new predictors of job performance (a half-day paper-and-pencil and computer test battery) also were collected during the CV.

The CV data collection procedures were subjected to extensive pilot and field tests. The data collection teams were extensively trained and were supervised by senior staff. The quality and completeness of the data collected attest to the thoroughness of these procedures. However, notwithstanding our best efforts, the final data were to some extent incomplete. The purpose of this paper is to describe the amount of missing data in the Project A

CV data base, the problems posed by incomplete data, and the steps taken to overcome those problems.

Reasons for Incomplete Data

Figure 1 lists some of the chief reasons for missing CV data. Most of the reasons are self-explanatory, but a couple examples involving the hands-on tests may help to illustrate some of the anticipated and unanticipated problems we encountered.

At Fort Hood, Texas, we were testing Armor Crewmen when a spring in the breech block of one of the howitzers failed. On that particular occasion, we had arranged for a back-up howitzer. Consequently, we did not lose any data.

During hands-on testing of Infantrymen at Fort Benning, Georgia, we were not so fortunate. The afternoon started bright and sunny. Consequently, we decided to administer the tests at our primary testing site, near a meandering creek, rather than at our back-up bad weather site. The weather in central Georgia is notoriously fickle on summer afternoons, though. A short time later, we were caught in a deluge. The creek rose. Everyone was up to their shins in water. And our test administrators were scrambling madly, trying to protect their equipment and score sheets from the driving rain. Unfortunately, one test administrator simply was not quick enough. As he and the hands-on test site manager watched, two scoresheets began to float away. A final effort was made to retrieve them, but before they could be reached, they were sucked into the creek and carried swiftly downstream. The thunderstorm abated a short while later. However, valuable time had been lost, and it was not possible to move all of the subjects through all of the test stations prior to the end of the soldier's work day. As a result, we were left with quite a bit of missing data.

Two other problems encountered during hands-on testing were equipment variation and score sheet errors. In most cases, we were able to make allowances for equipment variation by developing parallel forms of a test. Often, this involved omitting certain steps that were unnecessary for one of the equipment models. In other cases, parallel sets of steps were developed. We tried to make this clear on our score sheets and in scorer training. On a few occasions, we were unsuccessful. For example, in one hands-on test for Radio/Teletype Operators, the scoring sheets for one task included some steps to be scored for one type of equipment and a different set to be scored for another type of equipment. As a result, no subject should have had scores for all of the steps. Yet, two cases had data for both sets of steps, creating a unique problem of "too much" data rather than missing data. In several other instances, a scorer had trouble understanding some of the directions on the score sheets and left one or more steps unscored.

Figure 1

REASONS WHY DATA WERE SOMETIMES INCOMPLETE

HANDS-ON DATA

- Anticipated Variation in Equipment
- Unanticipated Variation in Equipment
- Soldiers Not Available for Part or All of Scheduled Time
- Equipment Breakdown or Nonavailability
- Conditions Preventing Testing of Some Soldiers on Some Tasks
- Scorer or Scoresheet Errors

RATING DATA

- No Suitable Raters Available
- Soldier Does Not Perform Some Kinds of Tasks
- Rater Not Following Instructions

KNOWLEDGE TEST

- Soldiers Not Available for Part or All of Scheduled Time
- Soldiers Exceptionally Slow in Taking Test
- Soldiers Not Following Instructions

Finally, a problem that plagued us throughout our testing was that subjects often had other commitments or were called away in the midst of tests. A subject might get halfway through a test, then have to leave for a dentist appointment that had been scheduled two or three months previously. These unavoidable absences doubtless caused more missing data than any other factor listed in Figure 1.

Amount of Missing Data

For any given instrument, data may be either partially missing (i.e., the soldier failed to complete some items or steps) or totally missing (i.e., the soldier was not available for a testing session). Moreover, if data are partially missing, there may be relatively small amounts of missing data or relatively large amounts of missing data.

Table 1 shows the extent of missing data for the school knowledge (SK) tests. There were only a very few instances (1%) where a soldier failed to take the test at all. There were also very few soldiers (1%) with relatively large amounts of missing data. There were, however, a significant number of cases (16%) with a small number of omitted items.

Table 1 also shows small differences between the Batch A and Batch Z MOS in the proportion of soldiers not tested at all. For all but one of the Batch A MOS, the percent not tested is above 1%, while the percent not tested is below 1% for all but one of the Batch Z MOS. This difference is a direct consequence of the fact that all of the Batch Z testing took place in a single day while the Batch A testing required two full days of a soldier's time.

Table 2 shows the extent of missing data for the job knowledge (JK) tests. (Subjects in Batch Z MOS did not receive job knowledge nor hands-on tests.) Again, there were very few instances (1%) where soldiers were not tested at all. The proportions of soldiers with relatively small (20%) and relatively large (3%) amounts of missing data are slightly higher than for the SK tests, but generally quite comparable.

The extent of missing data for the hands-on tests is shown in Table 3. The number of soldiers not tested was again small (1.8%). The number of soldiers with at least some missing data was, in many cases, very large. For the most part, this was due to equipment variation or failure.

Table 4 shows the extent of missing data for the rating measures. A scale or instrument was considered present if at least one peer or at least one supervisor provided a rating. With

TABLE 1

NUMBER AND PERCENT OF CASES WITH INCOMPLETE SR DATA FOR EACH MOS

MOS	No Missing	< 10% Miss	> 10% Miss	All Miss	Total
11B	604 86.04	88 12.54	2 0.28	8 1.14	702
13B	538 80.66	110 16.49	5 0.75	14 2.10	667
19E	403 80.12	88 17.50	4 0.80	8 1.59	503
31C	314 85.79	40 10.93	1 0.27	11 3.01	366
63B	536 84.14	81 12.72	10 1.57	10 1.57	637
64C	583 84.99	93 13.56	3 0.44	7 1.02	686
71L	458 88.11	41 7.98	2 0.39	13 2.53	514
91A	423 84.43	61 12.18	2 0.40	15 2.99	501
95B	583 84.25	100 14.45	3 0.43	6 0.87	692
12B	569 80.82	124 17.61	5 0.71	6 0.85	704
16S	402 85.53	67 14.26	0 0.00	1 0.21	470
27E	111 75.51	34 23.13	2 1.36	0 0.00	147
51B	88 81.48	14 12.96	5 4.63	1 0.93	108
54E	350 80.65	80 18.43	2 0.46	2 0.46	434
55B	209 71.82	65 22.34	15 5.15	2 0.69	291
67N	155 56.16	116 42.03	5 1.81	0 0.00	276
76W	388 79.18	90 18.37	10 2.04	2 0.41	490
76Y	487 77.30	119 18.89	19 3.02	5 0.79	630
94B	474 77.45	116 18.95	14 2.29	8 1.31	612
TOTAL	7675	1527	109	119	9430

TABLE 2

NUMBER AND PERCENT OF CASES WITH INCOMPLETE JK DATA FOR EACH BATCH A MOS

MOS	No Missing	< 10% Miss	> 10% Miss	All Miss	Total
11B	506 72.08	180 25.64	7 1.00	9 1.28	702
13B	460 68.97	180 26.99	17 2.55	10 1.50	667
19E	350 69.58	115 22.86	30 5.96	8 1.59	503
31C	304 83.06	24 6.56	31 8.47	7 1.91	366
63B	481 75.51	120 18.84	26 4.08	10 1.57	637
64C	533 77.70	141 20.55	5 0.73	7 1.02	686
71L	395 76.85	107 20.82	6 1.17	6 1.17	514
91A	428 85.43	59 11.78	9 1.80	5 1.00	501
95B	595 85.98	74 10.69	21 3.03	2 0.29	692
TOTAL	4052	1000	152	64	5268

TABLE 3

NUMBER AND PERCENT OF CASES WITH INCOMPLETE HANDS-ON DATA FOR EACH BATCH A MOS

MOS	No Missing	<10% Miss	>10% Miss	All Miss	Total
11B	188 26.78	471 67.09	30 4.27	13 1.85	702
13B	184 27.59	351 52.62	114 17.09	18 2.70	667
19E	341 67.79	131 26.04	18 3.58	13 2.58	503
31C	2 0.55	228 62.30	125 34.15	11 3.01	366
63B	135 21.19	380 59.65	106 16.64	16 2.51	637
64C	132 19.24	433 63.12	112 16.33	9 1.31	686
71L	244 47.47	218 42.41	46 8.95	6 1.17	514
91A	346 69.06	145 28.94	5 1.00	5 1.00	501
95B	326 47.11	308 44.51	56 8.09	2 0.29	692
TOTAL	1898	2665	612	93	5268

the exception of the Job Task Ratings (JTR), the completion rates were all quite high. The JTR scales provided a "cannot rate" option that was counted as missing, and this accounts for most instances of partially missing data. Tables 5 and 6 show the same information for supervisors and peers alone. The percent of soldiers with no ratings was quite a bit higher (8.4%) because no appropriate peer or supervisor was available in many instances.

Table 7 shows the amount of missing JTR (Batch A MOS) and CTR (Batch Z MOS) data by MOS. There was considerable variation across MOS. For some MOS (e.g., Combat Engineers, MANPADS Crewman) there were very high levels of completeness. However, for MOS where soldiers tend to work in isolation from other soldiers in their MOS and tend to perform only a subset of the tasks rated, the incidence of missing data was significantly higher. The best example is Administrative Specialist, where only 24% of the subjects had complete data.

From the results presented thus far, it might be tempting to conclude that, except for the JTR/CTR data, missing data was not a significant problem in analyses of the Project A CV data. Table 8 indicates that this is not the case. The table shows the number of Batch A soldiers with different patterns of complete and missing data across the four performance measurement methods. Fewer than 15% of the cases in the entire sample have complete data for all four methods. If the ratings data are set aside, there are still fewer than 25% of the subjects with complete HO, JK, and SK data. Similarly, ignoring the HO data still leaves only about 42% of the CV subjects with complete data on the remaining measures. Whether or not the sample of soldiers with complete data is representative of the target population, the sheer loss of statistical power associated with reduced sample size would be unacceptable. Something had to be done.

Treatment of Missing Data

The processing of missing data was approached in two stages. In the first stage, we focused on one instrument at a time and dealt with only those subjects who were missing a small amount of data on the instrument under consideration. In the second stage, we formulated procedures for dealing with subjects who were missing a high percentage or all of the data on a given instrument.

TABLE 4

PERCENT OF CASES WITH MISSING DATA BY RATING INSTRUMENT USING COMBINED SUPERVISOR AND PEER RATINGS (ALL MOS: N=9430)

Instrument	No Missing	1-10% Missing	> 10% Missing	All Missing
Army-Wide BARS	98.3	0.2	0.0	1.5
MOS Specific BARS	97.0	0.3	0.9	1.8
Task Ratings	66.2	11.2	20.1	2.4
Combat Prediction	98.3	0.1	0.1	1.5
All Instruments	66.0	28.7	3.8	1.5

TABLE 5

PERCENT OF CASES WITH MISSING DATA BY RATING INSTRUMENT FOR SUPERVISOR RATINGS ONLY (ALL MOS: N=9430)

Instrument	No Missing	1-10% Missing	> 10% Missing	All Missing
Army-Wide BARS	90.3	0.9	0.3	8.5
MOS Specific BARS	82.7	2.3	5.3	9.8
Task Ratings	30.2	13.5	45.3	10.9
Combat Prediction	89.4	1.8	0.2	8.6
All Instruments	29.2	50.0	12.3	8.4

TABLE 6

PERCENT OF CASES WITH MISSING DATA BY RATING INSTRUMENT FOR PEER RATINGS ONLY (ALL MOS: N=9430)

Instrument	No Missing	1-10% Missing	>10% Missing	All Missing
Army-Wide BARS	91.0	0.4	0.2	8.4
MOS Specific BARS	88.9	0.5	1.3	9.3
Task Ratings	48.1	11.0	30.0	11.0
Combat Prediction	90.4	0.9	0.2	8.6
All Instruments	47.5	34.2	9.9	8.4

TABLE 7

PERCENTAGE OF CASES WITH MISSING TASK RATINGS USING COMBINED PEER AND SUPERVISOR RATINGS

MOS	No Missing	1-10% Missing	> 10% Missing	All Missing	Total N
11B	71.51	14.39	12.39	1.71	702
13B	75.41	6.45	17.54	0.60	667
19E	68.79	14.51	16.30	0.40	503
31C	56.28	16.39	24.59	2.73	366
63B	63.27	10.99	22.92	2.83	637
64C	60.50	9.91	26.97	2.62	686
71L	23.93	18.29	53.89	3.89	514
91A	60.68	13.17	25.35	0.80	501
95B	70.38	11.85	17.63	0.14	692
12B	93.32	3.13	2.41	1.14	704
16S	91.49	5.32	3.19	0.00	470
27E	74.15	6.80	18.37	0.68	147
51B	84.26	5.56	8.33	1.85	108
54E	73.73	12.21	10.37	3.69	434
55B	69.42	12.71	16.15	1.72	291
67N	62.32	13.77	22.83	1.09	276
76W	61.22	14.49	20.20	4.08	490
76Y	49.05	11.59	31.43	7.94	630
94B	59.15	10.46	25.16	5.23	612
ALL MOS	66.18	11.20	20.22	2.40	9430

TABLE 8

NUMBER OF CASES WITH COMPLETE DATA FOR
EACH COMBINATION OF CRITERION INSTRUMENTS

BATCH A

Frequency Percent	Complete K3 & K5	Comp K3 Miss K5	Miss K3 Comp K5	Missing K3 & K5	TOTAL
Complete HO & RA	772 14.65	189 3.59	122 2.32	58 1.10	1141 21.66
Comp HO Miss RA	526 9.98	130 2.47	72 1.37	29 0.55	757 14.37
Miss HO Comp RA	1436 27.26	364 6.91	215 4.08	125 2.37	2140 40.62
Missing HO & RA	784 14.88	241 4.57	125 2.37	80 1.52	1230 23.35
TOTAL	3518 66.78	924 17.54	534 10.14	292 5.54	5268 100.00

Stage I: Missing Data within Each Instrument

Amount of missing data permitted. The first step in Stage I was to decide how much missing was too much. We examined distributions of the amount of missing data and found somewhat of a bimodal distribution. Most soldiers had only a small number of missing steps, items, or scales, but a smaller number had all or nearly all elements missing. For each instrument, we picked a percentage to be the dividing line between minimal and significant amounts of missing data. For cases with minimal missing data, we would take steps to fill in missing values so as to be able to compute performance scores. For cases with significant amounts of missing data, we would not attempt to compute performance scores for the instrument in question.

In general, we sought to retain 90 - 95% of the soldiers tested in each MOS, but to eliminate cases with more than 10% missing elements. For the written tests (JK and SK), we were able to set a 10% missing cutoff and still retain well over 95% of the subjects in each MOS. For HO and each of the rating instruments a slightly more liberal cutoff of 15% missing was chosen as the best balance between the desire to retain most of the cases and the desire to limit strongly the number of values that we would have to impute to achieve complete data. For the HO data a two-stage rule was adopted. For each task tested, we decided to generate a task score only if no more than 15% of the steps were missing. We then computed overall hands-on scores only if no more than 3 task scores (no more than 4 task scores for 31C and 63B, where we had relatively small samples) were missing.

Dropping unreliable responders. In addition to dropping cases with too much missing data on an instrument, we also developed rules for identifying and eliminating "unreliable" or random responders on each instrument. Again the rules were developed and adopted on an instrument-by-instrument basis. For the written tests, a random response index was defined as the correlation between the item score (1 for correct and 0 for incorrect) and item difficulty (expressed as proportion of subjects who answered the item correctly). For most examinees this correlation was positive since there was a tendency to get the easier items correct and miss the more difficult items. In a few instances this correlation was essentially zero, suggesting random responding. For these subjects, all of their responses for that particular instrument were set to missing

Random responding was not a concern for the hands-on data. The data sheets were filled out by trained (and monitored) NCOs and not by the examinees themselves. There was no indication that any subjects intentionally responded poorly or randomly in

front of the NCO scorers. No screening for unreliable responses in the hands-on data was conducted.

For the rating data, we screened for unreliable raters rather than unreliable examinees. We constructed reliability indices for each rater by comparing their ratings with the average of all other raters' ratings of the same soldiers on the same scales. Both mean difference and correlational indices were used in identifying "outliers" among the raters.

Establishing separate tracks to account for equipment differences. For several MOS, the hands-on scoring differed for different equipment. In order to achieve comparable scores across these equipment differences, we separated the examinees into separate "tracks" corresponding to the different variations in equipment. (For Military Police, for example, females use and were tested on a .38 caliber hand gun while males use and were tested on a .45 caliber hand gun.) We found at most minimal differences between track samples on those tasks that were scored the same, so we achieved comparable scoring by standardizing scores computed from tracked tasks separately for each track sample. Scores for each track were standardized to have a mean and standard deviation that matched the original overall mean for the score in question.

Number of subjects dropped for missing data or unreliable responses. The number of cases deleted due to too much missing data and/or to apparent random responding for the SK tests are shown in Table 9. Similar results for the JK tests are shown in Table 10 and the numbers of cases deleted due to too much missing data on the HO tests are shown in Table 11. Elimination of unreliable raters did not result in the loss of rating data for any individual subjects. In all cases, where raters were eliminated, there were other raters providing data on these subjects. (Where there were no other raters, the rater in question was not eliminated because there was no basis for estimating the reliability of the ratings.)

Imputing missing values. After dropping cases with too much missing data or with random responses, we imputed values for the remaining missing data so that summary performance scores could be computed.

We considered several options for imputing scores. The first was to compute the subject's mean on the variables that were present and then substitute this mean for each of the missing variables. If a subject passed 80% of the items or steps for a particular task or test, we could substitute a value of .8 for any missing item or step scores. This is equivalent to defining the total score as the mean of the values present, which was done in the field test. The problem with this approach was that items and steps differed considerably in difficulty. There were cases

TABLE 9

NUMBER OF CASES WITH SK DATA DELETED DUE TO TOO MUCH MISSING OR RANDOM RESPONSE

MOS	MISSING > 10%	RANDOM RESP	BOTH	TOTAL DROPPED	TOTAL N	PERCENT DROPPED
11B	2	8	0	10	694	1.4
13B-S	3	10	0	13	536	2.4
13B-T	2	2	0	4	117	3.4
19E	4	6	0	10	495	2.0
31C	1	5	0	6	355	1.7
63B	10	5	0	15	627	2.4
64C	3	7	0	10	679	1.5
71L	3	5	0	8	501	1.6
91A	2	5	0	7	486	1.4
95B	4	9	0	13	687	1.9
12B	5	11	0	16	698	2.3
16S	0	1	0	1	469	0.2
27E	2	3	0	5	147	3.4
51B	4	0	1	5	107	4.7
54E	2	4	0	6	432	1.4
55B	15	1	0	16	289	5.5
67N	5	0	0	5	276	1.8
76W	9	5	1	15	488	3.1
76Y	19	10	0	29	625	4.6
94B	14	20	0	34	604	5.6

TABLE 10

NUMBER OF CASES WITH JK DATA DELETED DUE TO TOO MUCH MISSING OR RANDOM RESPONSE

MOS	MISSING > 10%	RANDOM RESP	BOTH	TOTAL DROPPED	TOTAL N	PERCENT DROPPED
11B	9	6	0	15	693	2.2
13B	16	1	1	18	657	2.7
19E	29	6	1	36	495	7.3
31C	31	2	0	33	359	9.2
63B	26	4	1	31	627	4.9
64C	7	4	0	11	679	1.6
71L	6	1	0	7	508	1.4
91A	9	4	0	13	496	2.6
95B	22	3	0	25	690	3.6

TABLE 11

NUMBER OF CASES WITH HANDS-ON DATA DELETED DUE TO TOO MUCH MISSING

MOS	CASES DELETED	PERCENT DROPPED
11B	8	1.2
13B	37	5.7
19E	16	3.4
31C	14	4.1
63B	52	9.1
64C	37	5.5
71L	14	2.8
91A	0	0.0
95B	25	3.6

where the omitted items/steps were considerably more (or less) difficult than the items/steps that were completed, so systematic bias would be introduced by substituting the examinee's mean.

The second option that we considered was to substitute the variable (item, step, scale) mean for all missing values on that variable. This option was rejected because it would reduce individual differences. Subjects performing at different levels should have different estimates for the missing items.

The option used to fill in missing values was a procedure that had been developed for the National Center for Education Statistics (now the Center for Education Statistics) known as PROC IMPUTE (Wise & McLaughlin, 1980). Several features of PROC IMPUTE made it preferable to other readily available options for filling in the missing CV values.

First, PROC IMPUTE uses regression equations 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.

Second, PROC IMPUTE adds a random variable with variance equal to the error of estimate for predicting the missing value. If such a random variable is not added, the imputed values are more highly correlated with values on other variables in comparison with nonimputed values.

Third, PROC IMPUTE employs a sequential strategy that maintains relationships between variables when more than one value is imputed for the same examinee. A two-stage approach is employed so that the first variable is imputed from nonmissing values. The second (and subsequent) variable(s) are imputed from the nonmissing values plus the imputed value for the first variable. After all initial imputations, values are reimputed in a second pass where all of the initially imputed values participate in the reimputation of each missing value.

Finally, PROC IMPUTE models nonlinear relationships between the predicted and actual values. If the actual values are discrete, PROC IMPUTE provides discrete values for the missing elements as well. Table 12 illustrates the final step in PROC IMPUTE. The predicted values were divided into six equal intervals to define predicted "levels". There were from 61 to 92 cases at each predicted level for whom actual technical skill ratings were available. The distribution (in percentages) of actual scores for each predicted level is shown. For each soldier with a missing technical skill rating, a predicted level is computed. (Actually, the program interpolates between predicted levels.) A uniformly distributed random number between 0 and 100 is generated

and mapped onto the actual levels using the cummulative distribution of actual scores for the predicted level. (Again the program actually interpolates between levels.) The actual level scores are then transformed back to the original units.

Table 12

Distribution of Technical Skill Ratings for Each Predicted Level

Predicted Level	Total # of Cases	Percent at Each Actual Level				
		1	2	3	4	5
1	67	15	57	18	0	0
2	61	0	21	77	2	0
3	92	0	7	65	28	0
4	89	0	0	40	59	1
5	92	0	0	8	91	1
6	71	0	0	3	52	45

PROC IMPUTE was used in all instances except one. For the written tests, a distinction was made between internal omits (prior to the last item answered) and items that were not reached (omits after the last item answered). For internal omits, we assumed that the examinee did not know the answer and substituted a score equal to the guessing rate (e.g., .2 for a 5 option item). If the actual proportion passing the item was lower than the guessing rate, the proportion passing was used instead. We made no assumptions regarding items not reached since the examinee may not have had time to demonstrate knowledge of the item. Not reached items were imputed with PROC IMPUTE as were all missing hands-on steps and rating scales.

Tables 13, 14, and 15 show the changes in summary for statistics that resulted from the Stage I screening (pruning), standardizing (by track), and imputating for three different MOS. Initial totals were computed using means of available data. The sample sizes dropped slightly due to screening out random responders and cases with too much missing data. Only small changes in means, standard deviations, reliabilities and correlations resulted from the Stage I procedures. (Mean shifts for the first three scales should be compared against a standard deviation of 10.0, while the three rating factors were on a 7 point scale with a standard deviation of just under 1.0.)

Stage II: Missing Instruments

After cases were dropped or missing values were filled in on an instrument-by-instrument basis, we were ready to compute

Table 13

STAGE I RESULTS
CHANGES IN STATISTICAL CHARACTERISTICS OF SUMMARY PERFORMANCE MEASURES RESULTING FROM PRUNING, IMPUTING (EXCEPT RATINGS), AND STANDARDIZING
MOS 11B: INFANTRY

							CORRELATION WITH:			
PERF MEASURE	N	MEAN	SD	REL	SK	JK	HO	A1	A2	A3
SK TOTAL SCORE	-11	0.6	-0.7	-01	.	00	05	00	02	-01
JK TOTAL SCORE	-15	0.1	-0.3	01	00	.	05	01	01	00
HANDS-ON TOTAL	-4	-0.4	-0.9	02	05	05	.	05	02	06
AWB1: TECH/EFFORT	-1	-.01	.01	02	00	00	05	.	00	00
AWB2: INTEG&CONTR	-1	-.01	.01	02	02	01	02	00	.	01
AWB3: APPEARANCE	-1	.00	.01	02	-01	00	06	00	01	.

Table 14

STAGE I
CHANGES IN STATISTICAL CHARACTERISTICS OF SUMMARY PERFORMANCE MEASURES RESULTING FROM, IMPUTING (EXCEPT RATINGS), AND STANDARDIZING
MOS 63B: TRUCK MECHANIC

							CORRELATION WITH:			
PERF MEASURE	N	MEAN	SD	REL	SK	JK	HO	A1	A2	A3
SK TOTAL SCORE	-15	0.8	-0.7	-01	.	02	05	02	01	-03
JK TOTAL SCORE	-31	0.3	-0.4	00	02	.	08	00	01	-01
HANDS-ON TOTAL	-40	-0.3	-1.2	-03	05	08	.	-02	-01	-05
AWB1: TECH/EFFORT	-02	.01	.00	02	-02	00	-02	.	00	01
AWB2: INTEG&CONTR	-02	.01	.00	01	00	01	-01	00	.	00
AWB3: APPEARANCE	-02	.01	.00	02	-03	-01	-06	01	00	.

Table 15

STAGE I
CHANGES IN STATISTICAL CHARACTERISTICS OF SUMMARY PERFORMANCE MEASURES RESULTING FROM PRUNING, IMPUTING (EXCEPT RATINGS), AND STANDARDIZING
MOS 71L: CLERK

							CORRELATION WITH:			
PERF MEASURE	N	MEAN	SD	REL	SK	JK	HO	A1	A2	A3
SK TOTAL SCORE	-03	0.4	-0.3	-01	.	01	03	00	00	-01
JK TOTAL SCORE	-07	0.1	-0.1	00	01	.	02	01	01	00
HANDS-ON TOTAL	+06	-7.0	-0.8	05	03	02	.	00	00	-01
AWB1: TECH/EFFORT	00	-.00	.00	00	02	01	-01	.	00	00
AWB2: INTEG&CONTR	00	.00	.00	00	02	01	00	00	.	00
AWB3: APPEARANCE	00	.00	.00	00	-01	00	-01	00	00	.

Table 16

NUMBER OF CASES
MISSING EACH INSTRUMENT
(After Stage I Screening and Imputation)

	11B	13B	19E	31C	63B	64C	71L	91A	95B
Total N	702	667	503	366	637	686	514	501	692
Missing Hands-On	20	55	29	25	68	46	20	5	27
Missing Job Kn	24	29	44	40	41	18	13	18	29
Missing Scho Kn	18	28	18	17	25	17	21	22	13
Missing AW Bars	7	2	1	8	12	8	11	3	0
Missing MOS Bars	9	12	3	9	18	13	23	8	0
Missing Comb Pred	7	2	1	8	12	8	11	3	0
Missing A1: Awards	14	24	13	13	11	12	14	11	4
Missing A1: Phys Red	63	93	53	30	80	81	60	59	57
Missing A4: Arts. 15	23	28	16	14	11	14	15	14	4
Missing A5: Prom Rt	109	143	83	62	97	86	79	61	84
Total Complete	512	406	335	241	411	486	355	374	513
% Complete	72.9	60.9	66.6	65.9	64.5	70.9	69.1	74.7	74.1

Final Counts After Stage II Imputation

Total N	693	656	490	356	615	675	506	492	686
% of Original	98.7	98.4	97.4	97.3	96.6	98.4	98.4	98.2	99.1

Table 17

STAGE II
CHANGES IN STATISTICAL CHARACTERISTICS OF SUMMARY PERFORMANCE MEASURES RESULTING FROM STAGE II IMPUTATIONS
MOS 11B: INFANTRY

						CORRELATION WITH:			
PERF MEASURE	N	MEAN	SD	SK	JK	H0	A1	A2	A3
SK TOTAL SCORE	+11	-.1	+.1	.	01	-02	02	01	01
JK TOTAL SCORE	+16	-.5	+.1	01	.	00	01	00	01
HANDS-ON TOTAL	+15	-2.0	-.5	-02	00	.	02	01	03
AWB1: TECH/EFFORT	0	.0	.01	02	01	02	.	00	00
AWB2: INTEG&CONTR	+6	-.01	+.01	01	00	01	00	.	00
AWB3: APPEARANCE	+6	.00	.00	01	01	03	00	00	.

Table 18

STAGE II
CHANGES IN STATISTICAL CHARACTERISTICS OF SUMMARY PERFORMANCE MEASURES RESULTING FROM STAGE II IMPUTATIONS
MOS 63B: TRUCK MECHANIC

						CORRELATION WITH:			
PERF MEASURE	N	MEAN	SD	SK	JK	H0	A1	A2	A3
SK TOTAL SCORE	+13	-.4	-.0	.	00	01	01	00	00
JK TOTAL SCORE	+25	1.0	.0	00	.	02	01	00	02
HANDS-ON TOTAL	+49	-.9	-.7	01	02	.	00	00	05
AWB1: TECH/EFFORT	0	.00	.00	00	00	00	.	00	00
AWB2: INTEG&CONTR	+6	.00	.00	00	00	00	00	.	00
AWB3: APPEARANCE	+9	.00	.01	00	02	05	00	00	.

Table 19

STAGE II
CHANGES IN STATISTICAL CHARACTERISTICS OF SUMMARY PERFORMANCE MEASURES RESULTING FROM STAGE II IMPUTATIONS
MOS 71L: CLERK

						CORRELATION WITH:			
PERF MEASURE	N	MEAN	SD	SK	JK	H0	A1	A2	A3
SK TOTAL SCORE	+11	.0	.0	.	00	-03	00	00	00
JK TOTAL SCORE	+6	.8	.0	01	.	-01	00	-02	00
HANDS-ON TOTAL	+18	-4.3	-1.7	-03	-01	.	-01	00	02
AWB1: TECH/EFFORT	0	.00	.00	00	00	-01	.	00	00
AWB2: INTEG&CONTR	+7	.00	.00	00	-02	00	00	.	01
AWB3: APPEARANCE	9	.00	.00	00	00	-02	00	01	.

overall performance scores that combined information from the different measurement methods. The decision at this stage was whether to estimate individual scores if only partial data were available for the individual. We decided on a 50% rule. An examinee had to have data on at least half of the instruments going into a particular performance construct before we would estimate a score on the performance construct. Where 50% or fewer of the pieces were missing, PROC IMPUTE was again used to fill in the missing pieces.

Table 16 shows the number of soldiers in each MOS who had missing values for each instrument after the completion of the Stage I imputations and screening. In most instances, the number of missing cases was quite small (1 or 2%). The chief exceptions were two of the administrative measures. (Administrative measures were not included in stage I imputations because they do not include a large number of component parts.) Physical Readiness test scores were missing for 10 to 15% of the examinees. In most instances, peer and supervisor ratings of physical fitness were available for these same examinees. Similarly, Promotion Rate Deviation scores were missing for a significant number of cases (15%). This was primarily due to problems in retrieving Accession file information needed to compute time-in-service. For the most part, variation in promotion rates among first tour enlisted soldiers reflected
instances where disciplinary problems led to delays in promotions. Such delays were predicted fairly well from ratings of self control and integrity and from the administrative index of disciplinary actions.

Tables 17, 18, and 19 show changes in summary statistics that resluted from Stage II imputations for the same three MOS as before. Again only small changes resulted. There was a slight drop in hands-on means, because soldiers with missing hands-on scores tended to score well below average on other measures.

Summary

The decision rules and imputation procedures used with the CV data were successful in allowing us to develop performance scores for a very high proportion of the soldiers tested. Based on the available evidence, we have no reason to believe that any significant distortions were introduced while achieving this goal. Relatively few values were imputed at all. Where imputation was necessary, it was done with great care.

The apparent ease of imputation procedures should not, however, lead us to relax our data collection procedures in the future. Lessons learned from investigation of the reasons for missing data will be used to modify data collection procedures for the Project A longitudinal validation so as to further reduce the amount of missing data.

REFERENCE:

Wise, L. L. & McLaughlin, D. H. (1980). Guidebook for the imputation of missing data. Palo Alto, CA.: American Institutes for Research.

DEVELOPMENT OF PROJECT A JOB PERFORMANCE MEASURES

Charlotte Campbell
Human Resources Research Organization

Walter C. Borman
Personnel Decisions Research Institute

Daniel C. Felker
American Institutes for Research

Pat Ford
Maria De Vera Park
Human Resources Research Organization

Elaine C. Pulakos
Personnel Decisions Research Institute

Barry J. Riegelhaupt
Human Resources Research Organization

Michael G. Rumsey
U.S. Army Research Institute

Presented at the Annual Conference of the
Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

DEVELOPMENT OF PROJECT A JOB PERFORMANCE MEASURES

Charlotte H. Campbell, Walter C. Borman, Daniel C. Felker,
Pat Ford, Maria de Vera Park, Elaine C. Pulakos,
Barry J. Riegelhaupt, and Michael G. Rumsey

You have heard from the previous presenters about the overall objectives of Project A, and about the predictor development. The purpose of this paper is to describe the objectives, procedures, and products of the criterion development work.

The overall strategy for performance (i.e., criterion) measurement in Project A was to define the total domain of Army entry-level enlisted personnel performance in and then develop reliable and valid measures of all of the major components. The specific measures would be used as criteria against which to validate the predictor measures.

In defining the performance domain, we began with two assumptions. The first is that job performance is multidimensional. There is no one attribute, outcome, or factor that can be labeled as job performance. The second assumption is that job performance is manifested by a wide variety of behaviors or activities, things people do, that are judged to be important for accomplishing the goals of the organization. Each of these activities probably requires different knowledges and skills which are in turn most likely a function of different abilities.

For any particular job, one fundamental task in defining the performance domain is to describe the basic factors that comprise performance. For the population of entry-level positions in the Army, we postulated that there are two major types of job performance factors. The first is composed of performance components that are specific to a particular job, such as typing for the administrative specialists or loading the tank gun for tank crewmen; these we have labeled "job-specific" criterion factors. The second type of performance includes components that are defined and measured in the same way for every job (Borman, Motowidlo, Rose, & Hanser, 1985). These have been referred to as "Army-wide" criterion factors. Examples might be proficiency on the tasks for which every soldier is responsible, or demonstrating peer leadership or support.

The initial working model of total performance viewed performance as multidimensional within the job-specific and Army-wide factors. The job analysis and criterion construction methods were designed to "discover" the content of these factors via a comprehensive description of the total performance domain, several iterations of data collections, and the use of multiple methods for identifying and measuring the basic performance factors.

Defining the Job Content Domain

The definition of the job content domain was approached from several angles, including collection of critical incidents, review of Army job and task analyses, and review of Army training programs.

Critical incidents.

Through the conduct of critical incident workshops, Army personnel provided hundreds of critical incidents of specific task performance within each focal job, and thousands of critical incidents describing performance behaviors that have a general, not job-specific, referent. These large samples of job behaviors were translated into dimensions which identify both job-specific and Army-wide performance factors.

Army job and task analyses.

The Army maintains complex and definitive job and task analyses for every enlisted job. These include lists of the tasks required of all soldiers, regardless of their jobs, and provides step-by-step descriptions of how the task is to be performed, under what conditions, and to what standard. Another part of the system lists the tasks required for soldiers in the specific jobs and provides similar detailed task analyses. The Army Occupational Survey Programs are task inventories for specific jobs, which are administered periodically to soldiers in the jobs to determine which of hundreds of tasks and activities are performed by soldiers at various levels within the Army.

For each job, a data bank of task statements was accumulated from the integration of these sources, and the individual task statements were edited to determine if they indeed focused on observable job tasks, if they were redundant or overlapped with other tasks, if they were required only for soldiers in restricted or specialized assignments, and if they were at the same level of generality. Army job experts reviewed these edited lists to determine whether they provided a complete picture of the job requirements. The result was a task-based definition of the job performance domain.

Army training programs.

Prior to beginning work in any job, soldiers attend training courses that cover both basic Army soldiering skills and job-specific skills. As a matter of Army policy, training must be job-related; therefore examination of the training curricula should provide another view of the domain of job performance. Working with each of the schools where this training is developed and/or administered for the 19 jobs, we developed descriptions of the objectives and content of the training curricula. Job and task analyses, described above, were used in conjunction with these descriptions to develop detailed descriptions of training for each job. What was produced was a thorough analysis of the objectives, curriculum, and assessment procedures for the key schools.

Representing and Measuring the Job Content Domain

The criterion development work was guided by the desire to cover as many bases as possible relative to the population of criterion measures that it is possible to collect. We know a lot more about predictor constructs than we do about job performance constructs. There are volumes of research on the former, and almost none on the latter. For personnel psychologists it is almost second nature to talk about predictors in terms of constructs. However, investigation of job performance constructs seems limited to those few studies dealing with synthetic validity and those using the critical incidents format to develop performance factors. Relatively little attention has been given to conceptualizing performance in clerical, technical, or skilled jobs. Because we know so little about the underlying structure of job performance, we used every bit of measurement technology we had.

Our use of the technology was, we hoped, even-handed with respect to methods of defining performance and developing measures. We would be hard-pressed to defend placing the criterion variables on some continuum from immediate, through intermediate, to ultimate as a means for portraying their relative importance or functional interrelationships. For example, although there are good reasons for developing hands-on (job sample) performance measures, we would not be willing to defend hands-on performance scores as the "most ultimate" measure. And although job analyses based on critical incidents enjoy great respect and intuitive appeal, we would not propose these analyses as the "most valid" definitions of performance requirements.

Although our efforts involved intensive examination of the job-specific domain for all of the 19 jobs, intensive measurement was to be focused on nine of those jobs; fewer job-specific criterion measures were to be developed for the other ten jobs.

Analysis of the critical incidents led to the development of two sets of behaviorally anchored rating scales (BARS). One set, which was based on those behavioral examples which were tied to performance of job-specific activities, consists of six to twelve scales for each of the nine intensively-studied jobs. The other set was developed from the non-job-specific examples, and consists of 11 Army-wide scales, which would be used to assess performance of soldiers in all 19 jobs. On these and on all other rating scales, soldiers would be rated by peers and supervisors. Figure 1 presents an example of the job-specific dimensions for one of the nine jobs, and lists the Army-wide dimensions which applied to all jobs. It should be noted that what we developed were behavioral summary scales, containing anchors that represent the behavioral content of all performance incidents reliably retranslated by Army job experts for that particular level of effectiveness.

The task-based domain lists were clustered sampled by Army personnel with experience in each of the nine jobs, who also provided judgments of the importance of each task and the expected performance level and variability of each task for entry-level soldiers. Other job experts then sampled 30 tasks from each domain using the clusters and judgments. For each of the 30 tasks that they selected for each job, we developed multiple-choice paper-and-pencil job knowledge tests; for 15 of those tasks, we also constructed hands-on job sample tests. In several of the jobs, we developed parallel versions of the job knowledge and hands-on tests in order to cover various equipment systems in operation. Figure 3 lists the tasks for which tests were developed for one of the nine jobs.

The analysis of the Army's job training programs included grouping of the training objectives into duty areas, corresponding to the grouping of tasks in the occupational surveys. Multiple-choice paper-and-pencil tests of training achievement were constructed for the tasks in the duty areas, for each of the 19 jobs. In order to cover the "incidental learning" of job tasks not covered specifically during training, each test also included items for tasks not included in the curricula. Army job experts rated each item on its importance and relevance to training and to job performance; these ratings were used in selecting items to appear on the tests. Figure 4 presents the duty areas for which items were developed for one of the 19 jobs.

As we went through these development activities, we became aware of potential shortcomings in the set of performance measures, which could prove to be important in interpreting results. Accordingly, additional measures were developed. Three were to be administered for soldiers in all 19 jobs (Figure 5). These included a single rating scale of Overall Effectiveness, on which the rater was to consider performance in all of the categories on the Army-wide BARS instrument; a single rating of NCO (noncommissioned officer) Potential, which might well be independent of the Army-wide and Overall Effectiveness ratings; and a single rating scale of Overall Performance On Specific Job Duties.

Another area which was explored concerned records of administrative actions, which the Army routinely maintains for all enlisted personnel. Most of the information is maintained in noncomputerized files at the soldier's unit of assignment; some is also forwarded to central computerized files. Because obtaining the information from noncomputerized files was excessively labor-intensive, we developed a self-report form, which asked for information in five areas, including awards, letters of commendation, and disciplinary actions; these seemed, on the basis of their base rates and judged relevance, to have at least some potential for service as criterion measures. This Personnel File Information Form was also to be administered to soldiers in all 19 jobs.

For the ten jobs which were being studied less intensively, we developed a set of rating scales covering performance on 13 Common Task dimensions, the Common Tasks being those which are required of all soldiers, regardless of their jobs. The 13 dimensions include such things as basic first aid and firing of individual weapons (see Figure 6). These were the only source of job- or task-specific information which would be obtained from soldiers in the ten jobs.

For the nine jobs which were subjected to intensive study, sets of rating scales covering performance on the 15 tasks tested hands-on were constructed; although these were not behaviorally anchored scales, we hoped that they would provide a link between the job-specific BARS and the hands-on performance results. A Job History Questionnaire, requesting indication of the recency and frequency of performance on the tasks covered by the job knowledge and hands-on tests, was also developed, in order to assess the likely impact of experience effects on task performance. (Figure 7)

Army Management Agency Reviews

Throughout the initial development cycle, spanning the first three years of the project, we sought and received extensive involvement from the Army management agencies responsible for setting training and job performance policy. All of our procedures for obtaining information, all of our instruments, items, scales, and instructions, all of our data collection plans, were closely monitored by personnel from these agencies. By means of a series of formal briefings and reviews and informal discussions, we received valuable advice and direction concerning future planning, projections, and policies. Such management involvement has been and will continue to be invaluable in maintaining the integrity of the criterion development.

Pilot Testing and Field Tryouts

All of the measures went through several iterations of pilot testing and larger-scale field tryouts before they were finalized for the Concurrent Validation. The pilot tests were used to insure the technical accuracy, readability, and acceptability of the measures. The field tryouts served as a dry-run for the Concurrent Validation. They involved testing of 114 to 178 soldiers in each of the nine intensively studied jobs, using all of the Army-wide and job-specific instruments; tryouts of the training achievement tests among soldiers completing job training provided the needed information for the other ten jobs. Results were used to revise and refine the instruments. The training achievement tests, job knowledge tests, and expected combat performance scales were reduced in order to be administrable within the time allotted for the Concurrent Validation, a small number of scales were dropped, hands-on tests were revised to insure reliable observation and scoring, variables were added to the personnel information form, instructions were refined.

The final array of the criterion measurement instruments is portrayed in Figure 8. These were the tests and scales that were used in the Concurrent Validation. The next papers will describe how the data on those instruments and the predictor instruments were collected and analyzed.

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

FIGURE 1

ARMY-WIDE RATING SCALES (BARS) DIMENSIONS

- Technical knowledge/skill
- Initiative/effort
  .
  .
  .
- Self-control

JOB-SPECIFIC RATING SCALES (BARS) DIMENSIONS FOR CANNON CREWMEN

- Loading out equipment
- Driving and maintaining vehicles, howitzers, and equipment
  .
  .
  .
- Position improvement

FIGURE 2

COMBAT PERFORMANCE PREDICTION SCALE DIMENSIONS

- Cohesion/commitment to others
- Self-discipline/responsibility
- Mission orientation
- Technical/tactical knowledge
- Initiative

FIGURE 3

TRAINING ACHIEVEMENT TEST DUTY AREAS FOR CANNON CREWMAN

- Cannon equipment emplacement/displacement
- Firing battery operations during firing
  .
  .
  .
- Communications equipment and operator maintenance

TASKS COVERED BY JOB KNOWLEDGE AND HANDS-ON JOB SAMPLE TESTS FOR CANNON CREWMAN

- Perform cardiopulmonary resuscitation *
- Prevent shock
  .
  .
  .
- Disassemble/assemble breech (M109; M110; M198; M102) *

* Hands-on test developed (all tasks covered by job knowledge tests).

FIGURE 5

ADDITIONAL MEASURES FOR SOLDIERS IN ALL 19 JOBS

- Single scale rating of Overall Effectiveness
- Single scale rating of NCO Potential
- Single scale rating of Overall Performance on Specific Job Duties
- Personnel File Information Form - Variables:
  Number of awards and decorations
  Number of letters/certificates of appreciation, commendation, achievement
  Number of Articles 15/Flag actions (Disciplinary actions)
  Number of Military Training Courses

FIGURE 6

ADDITIONAL MEASURES FOR SOLDIERS IN TEN JOBS

- Rating scales on Common Task Areas:
  See: Identifying Threat (armored vehicles, aircraft)
  See: Estimating Range
  .
  .
  .
  Survive: Knowing and Applying the Customs and Laws of War

FIGURE 7

ADDITIONAL MEASURES FOR SOLDIERS IN NINE JOBS

- Task Performance Rating Scales - on the 15 tasks tested hands-on
- Job History Questionnaire - recency and frequency of performance on 30 tasks in job knowledge and/or hands-on tests

---

FIGURE 8

CRITERION MEASUREMENT INSTRUMENTS FOR CONCURRENT VALIDATION

Performance Measures For All 19 Jobs:

- Army-Wide Rating Scales (all obtained from both supervisors and peers).
  - Ten behaviorally anchored rating scales (BARS) designed to measure factors of non-job-specific performance.
  - Single scale rating of Overall Effectiveness.
  - Single scale rating of NCO Potential.
- Combat Performance Prediction Scale (obtained from both supervisors and peers) containing 40 items.
- Paper-and-pencil test of Training Achievement, developed for each of the 19 jobs (130-210 items each).
- Personnel File Information Form, developed to gather objective archival records data (awards and letters, rifle marksmanship scores, physical training scores, etc.). Self-report.

Performance Measures for Nine Jobs Only:

- Job Sample (Hands-On) tests of job-specific task proficiency.
  - Individual is tested on each of about 15 major job tasks in a job.
- Paper-and-pencil Job Knowledge Tests designed to measure task-specific job knowledge.
  - Individual is scored on 150 to 200 multiple-choice items representing about 30 major job tasks. Ten to 15 of the tasks were also measured hands-on.
- Rating scale measures of specific task performance on the tasks measured with the hands-on tests.
- Job-Specific Rating Scales (obtained from both supervisors and peers).
  - From 6 to 12 behaviorally anchored rating scales (BARS), developed for each job, to represent the major factors that constitute job-specific technical and task proficiency.
  - Single scale rating of Job Performance
- A Job History Questionnaire which asks for information about frequency and recency of performance of the job-specific tasks (self-report).

Performance Measures for Ten Jobs Only:

- Army-Wide Rating Scales (all obtained from both supervisors and peers).
  - Ratings of performance on common tasks (e.g., basic first aid).
  - Single scale rating on performance of specific job duties.

CREDITS

Borman, W. C., Motowidlo, S. J., Rose, S. J., & Hanser, L. M. (October, 1985) "Development of a Model of Soldier Effectiveness." ARI Technical Report.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (October, 1985) "Development and Field Test of Task-Based MOS-Specific Criterion Measures." ARI Technical Report 717.

Campbell, J. P. (Ed.) (October, 1985) "Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1985 Fiscal Year." ARI Technical Report.

Campbell, J. P., & Harris, J. H. (August, 1985) "Criterion Reduction and Combination via a Participative Decision-Making Panel." Paper presented at Annual Convention of the American Psychological Association, Los Angeles.

Davis, R. H., Davis, G. A., Joyner, J. N., & de Vera, M. V. (October, 1985) "Development and Field Test of Job-Relevant Knowledge tests for Selected MOS." ARI Technical Report.

Pulakos, E. D., & Borman, W. C. (Eds.) (October, 1985) "Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program." ARI Technical Report 716.

Riegelhaupt, B. J., Harris, C. D., & Sadacca, R. (October, 1985) "Development of Administrative Measures as Indicators of Soldier Effectiveness." ARI Technical Report.

Riegelhaupt, B. J., & Sadacca, R. (in preparation) "Development of Combat Prediction Scales." ARI Technical Report.

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (in preparation) "Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS." ARI Technical Report.

ANALYSIS OF CRITERION MEASURES: THE MODELING OF PERFORMANCE

John P. Campbell
Human Resources Research Organization

Jeffrey J. McHenry
Lauress L. Wise
American Institutes for Research

Presented at the Annual Conference of the
Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

The paradigm of Project A is simply that of a criterion related validity study, albeit a very large one that examines an entire system at once. Previous papers have discussed predictor development, criterion development, and data editing and preparation. This paper is intended to illustrate further the usefulness of a good theory, or even a poor one, in applied research. It recounts our attempt to model job performance in this population of jobs and to maximize our understanding of the previously described criterion measures. Recall that multiple methods were used to assess individuals on a wide array of performance components. Great care was taken in the task analysis and critical incident analysis to build in as much content validity as possible and considerable resources were devoted to careful measurement development.

THE INITIAL FRAMEWORK

The overall criterion development work was guided by a particular "theory" of performance, the basic outline of which is as follows. First, job performance really is multi-dimensional. There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization.

Two General Factors

For the population of entry level enlisted positions we postulated that there are two major types of job performance components. The first is composed of components that are specific to a particular job. That is, measures of such components would reflect specific technical competence or specific job behaviors that are not required for other jobs. The second kind of performance factor includes components that are defined and measured in the

same way for every job. These are referred to as Army-wide criterion factors.

For the job specific components, we anticipated that there would be a relatively small number of distinguishable factors of technical performance that would be a function of different abilities or skills and which would be reflected by different task content.

The Army-wide concept incorporates the basic notion that total performance is much more than task or technical proficiency. It might include such things as 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 the project began viewed performance as multi-dimensional within the two broad categories of factors. The job analysis and criterion construction methods were designed to "discover" the content of these factors via an exhaustive description of the total performance domain, several iterations of data collection, and the use of multiple methods for identifying basic performance factors.

Factors vs. a Composite

Saying that performance is multi-dimensional does not preclude using just one index of an individual's contributions to make a specific personnel decision (e.g., select/not select, promote/not promote). As argued by Schmidt and Kaplan (1971) some years ago, it seems quite reasonable for the organization to scale the importance of each major performance factor relative to a particular personnel decision that must be made and to combine the weighted factor scores into a composite that represents the total contribution or utility of an individual's performance, within the context of that decision. That is, the way in which performance information is weighted

and combined is a value judgment on the organization's part. The determination of the specific combinational rules (e.g., simple sum, weighted sum, non-linear combination) that best reflect what the organization is trying to accomplish is a matter for research.

Needed: The Latent Structure of Performance

If all the rating scales are used separately, the MOS-specific measures are aggregated at the task or instructional module level, and the major predictor subscales are used, there are approximately 200 criterion scores on each individual, which is too many to handle. Adding them all up into a composite is a bit too atheoretical and developing a reliable and homogeneous measure of the general factor violates the basic notion that performance is multi-dimensional. A more formal way to model performance is to think in terms of its latent structure, postulate what that might be and then resort to a confirmatory analysis. Unfortunately, it is true that we simply know a lot more about predictor constructs than we do about job performance constructs. There are volumes of research on the former, and almost none on the latter. For personnel psychologists it is almost second nature to talk about predictors in terms of theories and constructs. However, on the performance side, the textbooks are virtually silent. Only a few people have even raised the issue (e.g., Dunnette, 1963; Wallace, 1965).

Given this intial disparity, we used our own expert judgment, the previous literature, and data from pilot and field tests to formulate a target model. In the field tests, the various versions of the criterion measures were administered to 100-150 people from each of 9 MOS. These data and the development work leading up to them are summarized in Campbell (1985) and Campbell and Harris (1985). A picture we drew at the time is shown in Figure 1.

It is included only to show one stage in the almost continuous process of bootstrapping ourselves toward a more final conceptual description of the predictor/criterion space. The target model was then subjected to what might be described as a "quasi" confirmatory analysis using data from the concurrent validation sample. The purpose was to consider whether a single model of the latent structure of job performance would fit the data for all nine jobs. It is the results from these analyses that we report here.

PROCEDURE

As described previously, the final versions of the criterion measures were administered to a concurrent validation sample of 400-600 people in each of the 19 jobs (MOS). The complete array of performance measures is shown in Table 1.

The distinction between the Batch A (9 MOS) and Batch Z (10 MOS) is that not all criterion measures were developed for each job in Batch Z. Budget constraints dictated that the job-specific measures could only be developed for a limited number of jobs (i.e. Batch A).

Each "hands-on" test consisted of a number of critical steps, with each step scored pass or fail. The number of steps within a task varied widely from a half-dozen up to a maximum of 62. The job knowledge test consisted of 3 to 15 questions on each of a sample of 30 tasks (including the 15 also sampled for hands-on testing). The school knowledge test was organized around the "plan of instruction" in advanced individual (technical) training. Each test consisted of 100 to 200 items. The rating scales that were administered included 10 Army-wide (i.e. the scales were the same for all jobs) behaviorally anchored scales, from 8 to 13 job-specific behaviorally anchored

scales, ratings of performance on each of the 15 tasks tested hands-on, and a 40-item combat performance prediction questionnaire. Overall ratings of general effectiveness as a soldier potential for being an effective NCO were also obtained.

The performance indicators contained in official personnel records but obtained chiefly via self-report questionnaire, included such indicators as number of letters and certificates received, physical readiness test score, Articles 15 and other disciplinary actions, and M16 qualification level. File data were also used to construct a promotion rate score (relative to expected rate for a given length of service). The administrative measures were grouped into five scales on the basis of content; no attempts were made to further reduce these scales at this point.

RESULTS

The analysis had four major steps:

1. Determining a basic array of criterion scores that would constitute the input to the confirmatory analysis. In their unaggregated form, there were simply too many variables to theorize about.
2. Specification of a theory, or target matrix, that could be subjected to LISREL.
3. Determination of how much modification is necessary to fit the data adequately for each job.
4. Examining the fit of an overall model across all MOS.

Reduction of the Hands-On and Written Test Variables

Initial analyses indicated that individual task scores from the hands-

on and written job knowledge tests had only moderate internal consistency. Consequently, tasks were grouped by 6 research staff members into "functional or content categories" on the basis of similarity of task content. The 30 tasks sampled for each job were clustered into 8 to 15 categories. Each of the school knowledge items was similarly grouped into a specific content category.

Ten of the categories were common to some or all of the jobs (e.g., first aid, basic weapon, field techniques). Each job, except Infantryman, also had two to five performance categories that were unique or job specific. Figure 2 shows both the common and job specific item categories used for each of the nine jobs. Figure 3 includes a sample of the definitions that were generated for each content category.

Next, scores were computed for each content category within each of the three sets of measures. For the hands-on test, the functional category score was the mean percent of successfully completed steps across all of the tasks assigned to that category. For the job knowledge test and the school knowledge test, the functional category score was the percent of items within that category that were answered correctly.

After category scores were computed, they were factor analyzed via principal components. Separate factor analyses were executed for each type of measure within each job. There were several common features in the results. First, the unique or specific categories for each job tended to load on different factors than the common categories. Second, the factors that emerged from the common categories tended to be fairly similar across the nine different jobs and across the three methods. Some of the categories

were not sampled in one or more of the tests for some jobs, so some differences were inevitable.

Using these exploratory empirical factor analyses as a guide, the following set of content categories was identified.

1. Basic Soldiering Skills (field techniques, weapons, navigate, customs and laws).
2. Safety/Survival (first aid, nuclear-biological-chemical safety).
3. Communications (radio operation).
4. Vehicle Maintenance.
5. Identify Friendly/Enemy Aircraft and Vehicles.
6. Technical Skills (specific to the job).

At this point, the categories reflected an integration of expert judgment and the results of the factor analyses.

Reduction of the Rating Variables

As noted in a previous paper (Campbell, 1986), the individual ratings scales were reasonably reliable; however, the different scales exhibited intercorrelations varying from moderate to high. Further reduction in the number of scales was aimed at reducing redundancy and colinearity.

As also noted in a previous paper (Campbell, 1986), empirical factor analyses of the Army-wide rating scales suggested three factors. These were:

1. Effort/Leadership; including effort and competence in performing job tasks; leadership; and self-development.
2. Maintaining Personal Discipline: including self-control; integrity; and following regulations.

3. Fitness and Appearance: including physical fitness and maintaining proper military bearing and appearance.

Similar exploratory factor analyses were conducted for the job-specific BARS scales and two factors within each job were identified. The first consisted of scales reflecting performance that seemed to be most central to the specific technical content of each job. The second factor included the rating scales that seemed to reflect more tangential or less central performance components. Again the final formulation of factors was based on a combination of empirical and judgmental considerations.

The reliability, intercorrelations, and distributional properties of the task specific for each of the 30 tasks also tested with the knowledge tests were also examined. In general, these scales were less reliable than either the Army-wide or the job-specific behavioral summary scales. Supervisors and peers often reported that they had never had an opportunity to observe their ratees' performance on many of the tasks, leading to a significant missing data problem. Consequently, the task ratings were dropped from the present analyses.

The individual items in the combat performance prediction battery also were subjected to a principal components analysis. Two factors seemed to emerge from an analysis on the combined sample. The first factor consisted of items depicting exemplary effort, skill, or dependability under stressful conditions. The second factor consisted of items portraying failure to follow instructions and lack of discipline under stressful conditions.

The Final Array

Based on the above exploratory analyses, the reduced array of criterion variables for each job consisted of:

- 2-5 hands-on content category scores
- 2-6 job knowledge content category scores
- 2-6 school knowledge content category scores
- 3 Army-wide rating factors
- 2 job-specific rating factors
- 2 combat performance prediction rating factors
- 1 overall effectiveness rating
- 5 administrative measures scale scores

Tables 2 through 10 show the means, standard deviations, and intercorrelations among these variables for each of the nine jobs.

Building the Target Model

The next step was to build a target model of job performance that could be tested for goodness of fit within each of the nine jobs. The initial model shown in Figure 1 was a starting point. The correlation matrices shown in Tables 2 through 10 were each subjected to another round of empirical factor analysis to suggest possible modifications.

Several consistent results were observed in the different factor analyses. First, as expected, there was the general prominence of "method" factors, specifically one methods factor for the ratings and one methods factor for the written tests. The emergence of method factors was anticipated and was consistent with prior findings (e.g., Landy and Farr, 1980).

The second consistent result was a correspondence between the administrative measures scales and the three Army-wide rating factors. The awards

and certificates scale from the administrative measures loaded together with the Army-wide effort/leadership rating factor; the Article 15 and promotion rate scale loaded with the personal discipline factor (most of the variance in promotion rate was thought to be due to retarded advancement associated with disciplinary problems); and the physical readiness scale loaded with the fitness/appearance factor.

A third observation from the empirical factor analyses was that, with the possible exception of the job specific content factors, there was not much evidence that the factors reflecting task performance crossed measurement methods. The hands-on communication score, for example, was likely to be as correlated with the written safety score as with the written communication score. This result was taken as evidence against being able to separate measurement of task knowledge versus task performance skill within the common task domain.

Based on these findings from the exploratory empirical analyses, a revised model was constructed to account for the correlations among our performance measures. This model included the five job performance constructs shown in Figure 4. In addition, a "paper-and-pencil test" methods factor and a ratings "method" factor were retained.

Several minor issues remained before the model could be tested for goodness of fit within the nine Batch A jobs. One was whether the job-specific BARS rating scales were measuring job-specific technical knowledge and skill, or effort and leadership, or both. The intercorrelations among our performance factors suggested that these rating scales were measuring both of these performance constructs, though they seemed to correlate more highly with other

measures of effort and leaderhsip than with measures of job-specific technical knowledge and skill.

Another issue was whether it was necessary to posit hands-on and administrative measures "method" factors to account for the intercorrelations within each of these sets of measures. The average intercorrelation among the scores within each of these sets was not particularly high. Therefore, for the sake of parsimony, we decided to try to fit a model without these two additional methods factors.

Confirming the Model Within Each Job

The next step in the analysis was to conduct separate tests of goodness of fit of this target model within each of the nine jobs. This was done using the LISREL confirmatory factor analysis program (Joreskog & Sorbom, 1981).

In conducting a confirmatory factor analysis with LISREL, it is necessary to specify the structure of three different parameter matrices: Lambda-Y, the hypothesized factor structure matrix (a matrix of regression coefficients for predicting the observed variables from the underlying latent constructs); Theta-Epsilon, the matrix of uniqueness or error components (and intercorrelations); and Psi, the matrix of covariances among the factors. In these analyses, we set the diagonal elements of Psi (i.e. the factor variances) to one, forcing a "standardized" solution. This meant that the off-diagonal elements in Psi would represent the correlations among and between our performance constructs and method factors. We further specified that the correlation among the two method factors and each performance construct should be zero. This effectively defined the method factor as that portion of the common variance among measures from the same method that was not predictable from (i.e. correlated with) any of the other related factor or performance construct scores.

Some problems were encountered in fitting the hypothesized model for several of the jobs. Solutions were obtained with some factor loadings greater than one and with negative uniqueness estimates for the corresponding observed variables. Also, estimates of the correlations among the performance constructs occasionally exceeded unity. These problems necessitated a certain amount of ad hoc cutting and fitting in the form of computing the squared multiple correlation (SMC) for predicting each observed variable from all of the other variables, and setting the uniqueness estimates (i.e. Theta-Epsilon diagonal) to one minus this SMC. This approach eliminated all factor loadings and correlations greater than one. In most cases, a second "iteration" was performed to adjust the initial Theta-Epsilon estimates so that the diagonal of the estimated correlation matrix would be as close to one as possible.

Table 11 shows the final factor loading estimates from Lambda-Y for each job. Tables 12 and 13 show the uniqueness estimates from Theta-Epsilon and the factor intercorrelation estimates from Psi, respectively.

LISREL also computes a goodness-of-fit index based on a comparison of the actual correlations among the observed variables and the correlations estimated from Lambda-Y, Theta-Epsilon, and Psi. The goodness of fit is distributed as chi-square, with degrees of freedom dependent on the number of observed variables and the number of parameters estimated. The expected value of chi-square is equal to the degrees of freedom, it is a sign that the model does not fit the correlations among the observed variables.

Table 14 shows the value of chi-square for each job. These chi-square values should be interpreted with considerable caution. The approach we

used was not purely confirmatory. The hypothesized target model was based in part on analyses of these same data. In addition, LISREL was "told" that the Theta-Epsilon (uniqueness) parameters were all fixed, and therefore did not "use up" any degrees of freedom estimating these parameters; in fact, these values were estimated entirely from the data.

Confirmation of the Overall Model

The results of the above procedures applied to each job generally supported a common structure for job performance. The procedures also yielded reasonably similar estimates of the intercorrelations among the constructs and of the loadings of the observed variables on these constructs across the nine jobs.

The results of the confirmatory procedures applied to the performance measures from each job generally supported a common structure of job performance. The procedures also yielded reasonably similar estimates of the intercorrelations among the constructs and of the loadings of the observed variables on these constructs across the nine jobs.

The final step was to determine whether the variation in some of these parameters across jobs could be attributed to sampling variation. The specific model that we explored stated that (1) the correaltion among factors was invariant across jobs and (2) the loadings of all of the Army-wide measures on the performance constructs and on the rating method factor were also constant across jobs.

The proposed overall model was a relatively stringent test of a common latent structure. For example, it was quite possible that selectivity dif-

ferences in the different jobs would lead to differences in the apparent measurement precision of the common instruments or differences in the correlations between the constructs. This would tend to make it appear that the different jobs required different performance models, when in fact they do not.

The LISREL multi-groups option requires that the number of observed variables be the same for each job. However, virtually every job was missing scores on at least one of the five construct categories for at least one of the three knowledge and skill measurement methods. To handle this problem, the Theta-Epsilon error estimates for these variables were set at 1.00, and the observed correlations between these variables and all the other variables were set to zero. It was thus necessary to count the number of "observed" correlations that we generated in this manner and subtract this number from the degrees of freedom when determining the significance of the chi-square goodness-of-fit statistic.

The overall model fit extremely well. The root mean square residual was .047, and the chi-square was 2508.1. There were 2403 degrees of freedom after adjusting for missing variables and the use of the data in estimating uniquenesses. This yields a significance level of .07, not enough to reject the model. Tables 15 and 16 show the factor loadings and uniqueness for each job under this constrained model. Table 17 shows the final mapping of the criterion measures on the five performance factors.

Criterion Intercorrelations

Five residual scores were then created from the five criterion factors

in the following manner. A paper-and-pencil "methods" factor score was created by first summing the two paper-and-pencil knowledge tests (job knowledge and training content knowledge scores) and then partialing out the variance due to the correlation of the total paper-and-pencil test score with all non paper-and-pencil criterion measures (e.g., hands-on scores, rating scores, and administrative records scores). This residual was defined as the paper-and-pencil method score. This variable was in turn partialed from the Core Technical Proficiency criterion factor and from the General Task Proficiency factor creating two residual scores. A similar procedure was used to create a rating method factor score which was in turn partialed from the Effort/Leadership, Personal Discipline, and Physical Fitness/Military Bearing factors, thereby creating three more residual scores.

The five criterion factor scores, the five residual criterion scores, the single rating obtained from the overall performance rating scales, and the total score from the hands-on test were used to generate a 12 x 12 matrix of criterion intercorrelation for each MOS in Batch A. The averages of these correlations across MOS are shown in Table 18.

Remember that to create the residual scores the paper-and-pencil factor was partialed from the first two criterion factors and the rating method factor was partialed from the last three criterion factors. The intercorrelations of the 5 criterion factors are in the upper left quadrant, the intercorrelations among the 5 residual scores are in the lower right quadrant, and the cross correlations are in the upper right and lower left. Also remember that the first two factors contain items from both the knowledge tests and hands-on tests and the last three factors all contain both ratings and

administrative measures.

Some noteworthy features of this 12 x 12 matrix are the following.

- The intercorrelations of the factor pairs which confound measurement method (e.g., 1 with 2 or 3 with 4) are higher, as expected, than factor pairs which do not confound method (e.g, 1 with 3 or 2 with 4). However, they are not so high that collapsing the five factors into some smaller number would be justified. In fact, as illustrated later (McHenry), factors 1 and 2, which intercorrelate .531 on the average, yield different profiles of correlations with the selection tests.

- The correlation of the overall performance rating scale with the total hands-on test score is low (.203) but it is certainly not zero. Assuming a reliability of about .60 for each measure would yield an intercorrelation of about .34 when corrected for attenuation. Consequently, there is a substantial proportion of common variance between the two measures but by no means do they assess the same things. Assuming for the moment that the reliable variance in each measure is relevant to performance, a reasonable conclusion is that while performance on a standardized job sample is a significant component of performance it is by no means all of it.

- The correlations of the residualized factor 3 (effort/leadership residual) with the core technical factor, the re-

sidual core technical factor, the general task proficiency factor, the overall rating scale, and the hands-on total score are all about the same. Also, as compared to the correlation of the effort/leadership raw scores with these same variables, the correlations of effort/leadership residual with the core technical and general task proficiency factors go up while the correlations with personal discipline and physical fitness go down. Residualizing factor three (by removing the rating method factor) makes it more like a "can do" factor and less like a "will do" factor.

In general, these intercorrelations seem to behave in very lawful ways and are consistent with a multi-dimensional model of performance.

SUMMARY AND DISCUSSION

Several aspects of the final structure are noteworthy. First, in spite of some confounding factor content with measurement method, the latent performance structure appears to be composed of very distinct components. It is reaonsable to expect that the different performance constructs would be predicted by different things, so that validity generalization may not exist across the performance constructs within a job. If this is so, there is a genuine question of how the performance constructs should be weighted in forming an overall appraisal of performance for use in personnel decisions.

It is tempting to infer that Effort/Leadership and Maintaining Personal Discipline, particularly the latter, reflect aspects of performance that are under motivational control and consequently may be better predicted by personality or interest measures than by measures of ability or skill. This

leads us to the question of whether choices such as showing up on time, staying out of trouble, and expending extra effort under adverse conditions are a function of state or trait variables. We do have considerble data to focus on the question. It is also interesting that the residual score for factor 3 becomes more like a "can do" component of performance. It may be the case that raters cannot separate can do from will do when they are asked to retrospectively aggregate an individual's task performance and provide an evaluation of it. If the degree to which an individual exhibits a characteristic.effort level and consistency of performance is not task specific then halo might indeed be substantive variance and not error.

Given the high degree of consistency across jobs in the structure of the performance measures, it is worth asking to what extent our performance model generalizes to even wider domains of jobs. Some limitations appear likely. The "general soldiering skills" constructs would almost surely be quite different outside the military. Perhaps it would be replaced by a more generalized job skill construct. Similarly, it is likely that the physical fitness and military appearance construct also would be somewhat different for civilian occupations. The remaining constructs --technical skill, effort and leadership, and personal discipline-- all appear to be basic components of almost any job.

In generalizing to a wider domain of jobs, it is reasonable to suppose that other latent structures would fit other "populations" of jobs. For example, jobs that are not organized into units and that involve a great deal of written or oral communication (e.g., sales jobs) might have a different structure. It is tempting to ask how many different performance dimension structures define different populations of jobs. Such questions go well beyond the present finding, however, which is that a single structure <u>did</u> fit

the jobs studied.

Since (a) the five-factor solution is stable across jobs sampled from this population, (b) the performance constructs seek to make sense, and (c) the constructs are based on measures carefully developed to be content valid, it seems safe to ascribe some degree of construct validity to them.

REFERENCES

Campbell, J. P. (1986). When the textbook goes operational. Paper presented at the 94th Annual Convention of the American Psychological Association, Washington, D.C.

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

Dunnette, M. D. (1963). A modified model for selection research. Journal of Applied Psychology, 47, 317-323.

Joreskog, K. C., & Sorbom, D. (1981). LISREL VI: Analysis of Linear squares methods. Uppsala, Sweden: University of Uppsala.

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

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology, 24, 419-434.

Wallace, S. R. (1965). Criteria for what? American Psychologist, 20, 411-418.

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

Table 1

Summary of Criterion Measures Used in Concurrent Validation Samples[1]

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 (non-commissioned officer) potential.

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

5. A 40-item summated rating scale for the assessment of expected combat performance.

6. Five performance indicators from administrative records. The first four are obtained via self-report and the last one from computerized records.

   - o Total number of awards and letters of commendation.
   - o Physical fitness qualification.
   - o Number of disciplinary infractions.
   - o Rifle marksmanship qualification score.
   - o Promotion rate (in deviation units).

Performance Measures for Batch A Only

7. Job-sample (hands-on) test of MOS-specific task proficiency.

   - o Individual is tested on each of 15 major job tasks.

8. Paper-and-pencil job knowledge tests designed to measure task-specific job knowledge.

   - o Individual is scored on 150-200 multiple choice items representing 30 major job tasks. Fifteen of the tasks were also measured hands-on.

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

10. MOS-specific behaviorally anchored ratings scales. From 7 to 13 BARS were developed for each MOS to represent the major factors that constituted job-specific technical and task proficiency.

Performance Measures for Batch Z Only

11. 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.

Auxiliary Measures Included in Criterion Battery

12. Job History Questionnaire which asks for information about frequency and recency of performance of the MOS-specific tasks.

13. Work Environment Description Questionnaire - a 141-item questionnaire assessing situational/environmental characteristics, leadership climate, and reward preferences.

[1]All rating measures were obtained from approximately 2 supervisors and 3 peers for each ratee.

Table 2

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 11B: INFANTRY

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	Overall Rating	4.60	0.90	.	90	74	68	77	85	75	65	23	12	17	35	36	26	14	4	35	25	11	10	23	19	18	12	14
2	EFF/Ldr Rating	4.41	0.82	90	.	74	65	80	88	80	67	24	8	13	30	36	30	12	5	36	27	10	13	33	20	20	9	17
3	Discipline Rtng	4.50	0.87	74	74	.	49	55	71	63	66	13	3	7	39	31	16	10	3	30	22	6	8	24	13	13	5	13
4	Fitness Rating	4.86	0.89	68	65	49	.	59	66	52	45	17	27	9	24	22	10	9	-1	10	10	-2	-4	13	6	6	1	1
5	Job-Spec Tech	32.98	4.58	77	80	55	59	.	86	75	58	23	15	17	20	22	27	15	5	35	22	12	10	36	21	23	9	16
6	Job-Spec Other	22.67	3.66	85	88	71	66	86	.	80	67	25	8	14	28	32	23	10	6	35	26	12	12	33	17	22	11	17
7	Combat Exmplry	9.02	1.49	75	80	63	52	75	80	.	75	24	8	13	31	29	28	12	7	37	25	9	16	34	22	23	9	19
8	Combat Problems	10.03	1.64	65	67	66	45	58	67	75	.	14	8	6	33	27	20	7	-1	36	24	9	15	31	21	18	8	14
9	Awards & Certs	3.33	2.18	23	24	13	17	23	25	24	14	.	15	20	-2	4	13	6	-1	14	15	-0	13	9	9	5	4	12
10	Phys. Readiness	273.44	28.00	12	8	3	27	15	8	8	8	15	.	11	2	-6	1	-7	-9	0	5	-7	-0	8	-2	-1	-4	-6
11	M16 Qualifie.	2.74	0.57	17	13	7	9	17	14	13	6	20	11	.	1	1	13	6	-0	10	2	3	0	14	10	5	3	6
12	Articles 15	0.39	0.85	-35	-30	-39	-24	-20	-28	-31	-33	-2	2	1	.	-45	-10	-1	-6	-10	-9	-8	-6	-10	-1	-9	0	-5
13	Promotion Rate	0.03	0.68	36	36	31	22	22	32	29	27	4	-6	1	45	.	16	7	7	19	17	12	10	18	14	12	11	17
14	MO Basic	50.50	10.06	26	30	16	10	27	23	28	20	13	1	13	10	16	.	15	6	44	30	13	27	40	24	20	16	30
15	MO Safety	22.67	3.41	14	12	10	9	15	10	12	7	6	-7	6	-1	7	15	.	2	16	8	1	8	16	7	3	3	4
16	MO Comm	13.15	1.53	4	5	3	-1	5	6	7	-1	-1	-9	-0	-6	7	6	2	.	4	6	-1	-3	0	4	6	2	-1
17	JK Basic	50.93	9.71	35	36	30	10	25	35	37	36	14	0	10	10	19	44	16	4	.	68	40	42	65	50	40	30	25
18	JK Safety	20.02	4.31	25	27	22	10	22	26	25	24	15	5	2	9	17	30	8	6	68	.	23	26	47	41	32	25	20
19	JK Comm	4.37	1.47	11	10	6	-2	12	12	9	9	-0	-7	3	-8	12	13	1	-1	40	23	.	16	26	25	19	14	16
20	JK Identify	8.25	2.24	10	13	8	-4	10	12	16	15	13	-0	0	-6	10	27	8	-3	42	26	16	.	31	24	18	16	37
21	SK Basic	72.87	14.89	33	33	24	13	36	33	34	31	9	8	14	10	18	40	16	0	65	47	26	31	.	63	60	44	43
22	SK Safety	9.51	2.12	19	20	13	6	21	17	22	21	9	-2	10	-1	14	24	7	4	50	41	25	24	63	.	45	34	26
23	SK Comm	5.68	1.67	18	20	13	6	23	22	23	18	5	-1	5	-9	12	20	3	6	40	32	19	18	60	45	.	40	31
24	SK Vehicle	0.78	0.42	12	9	5	1	9	11	9	8	4	-4	3	0	11	16	3	2	30	25	14	16	44	34	40	.	21
25	SK Identify	2.80	1.16	14	17	13	1	16	17	19	14	12	-6	6	-5	17	30	4	-1	25	20	13	37	43	26	31	21	.

N= 503

Table 3

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 13B: CANNON CREWMAN

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	Overall Rating	4.59	0.79	.	86	71	61	62	72	73	61	11	10	5	-25	30	20	19	17	6	26	18	14	8	3	24	15	12	8	9
2	Eff/Ldr Rating	4.43	0.76	86	.	75	62	65	74	78	61	14	6	1	-23	25	27	25	14	9	32	20	15	11	5	30	20	15	5	13
3	Discipline Rtng	4.61	0.78	71	75	.	51	53	60	63	60	-0	-4	-1	-20	26	12	9	12	4	22	16	13	4	3	18	14	14	6	16
4	Fitness Rating	4.95	0.82	61	62	51	.	47	53	51	39	7	23	-1	-25	16	8	4	0	3	5	-1	-1	1	-2	4	-4	-1	-4	-6
5	Job-Spec Tech	23.39	3.35	62	65	53	47	.	80	60	39	11	10	1	-2	10	35	18	9	-1	25	10	10	17	8	24	8	12	6	4
6	Job-Spec Other	23.90	3.08	72	74	60	53	80	.	66	49	6	5	-4	-9	18	25	18	8	1	29	18	15	13	6	26	14	16	4	8
7	Combat Exmplry	9.00	1.44	73	78	63	51	60	66	.	63	14	10	3	-15	23	20	23	13	3	22	16	13	6	8	23	12	7	-1	1
8	Combat Problems	9.92	1.36	61	61	60	39	39	49	63	.	8	7	-3	-16	26	14	16	6	12	19	17	10	14	8	15	14	9	5	8
9	Awards & Certs	2.58	1.82	11	14	-0	7	11	6	14	8	.	12	18	0	8	15	19	15	-1	11	10	6	5	8	11	6	5	8	2
10	Phys. Readiness	261.74	32.70	10	6	-4	23	10	5	10	7	12	.	11	-3	-2	7	2	-7	8	-8	-8	-10	5	4	-0	-8	-10	-12	-15
11	M16 Qualific.	2.25	0.69	5	1	-1	-1	1	-4	3	-3	18	11	.	6	1	7	8	12	-3	-4	-5	-6	7	-3	-3	-7	0	3	-3
12	Articles 15	0.46	1.03	-25	-23	-20	-25	-2	-9	-15	-16	0	-3	6	.	-31	-0	-4	-5	-5	-7	-10	-12	-7	1	-5	-6	-2	-5	-3
13	Promotion Rate	0.01	0.63	30	25	26	16	10	18	23	26	8	-2	1	-31	.	6	10	10	3	10	6	5	5	-1	2	5	-2	10	7
14	MO Tech.	50.71	9.94	20	27	12	8	35	25	20	14	15	7	7	-0	6	.	47	20	11	33	13	7	10	12	36	18	20	11	9
15	MO Basic	48.50	13.00	19	25	9	4	18	18	23	16	19	2	8	-4	10	47	.	21	8	42	28	20	9	15	40	25	17	15	9
16	MO Safety	40.16	6.23	17	14	12	0	9	8	13	8	15	-7	12	-5	10	20	21	.	11	24	14	11	9	3	25	20	18	11	24
17	MO Comm	10.60	1.59	6	9	4	3	-1	1	3	12	-1	8	-3	-5	3	11	8	11	.	1	1	-2	6	5	7	5	-1	1	3
18	JK Tech.	50.67	9.94	26	32	22	5	25	29	22	19	11	-3	-4	-7	10	33	42	24	1	.	58	54	31	20	64	52	41	37	33
19	JK Basic	31.91	5.78	18	20	16	-1	10	18	16	17	10	-8	-5	-10	6	13	38	14	1	58	.	55	14	23	52	49	38	28	27
20	JK Safety	23.58	4.43	14	15	15	-1	10	15	13	10	6	-10	-6	-12	5	7	20	11	-2	54	55	.	10	21	41	38	35	26	27
21	JK Comm	1.12	0.68	8	11	4	1	17	13	6	14	5	5	7	-7	5	10	9	9	6	21	14	10	.	13	19	13	16	14	11
22	JK Identify	7.12	2.25	3	5	3	-2	8	6	8	8	8	4	-3	1	-1	12	15	3	5	20	23	21	13	.	20	21	25	10	9
23	SK Tech.	50.82	9.84	24	30	18	4	24	26	23	15	11	-0	-3	-5	2	36	40	25	7	64	52	41	19	20	.	63	47	38	40
24	SK Basic	23.17	5.27	15	20	14	-4	8	14	12	14	6	-8	-7	-6	5	18	25	20	5	52	49	38	13	21	63	.	51	40	52
25	SK Safety	8.44	2.12	12	15	14	-1	12	16	7	9	5	-10	0	-2	-2	20	17	18	-1	41	38	25	16	25	47	51	.	28	36
26	SK Comm	3.55	1.21	8	5	6	-4	6	4	-1	5	8	-12	3	-5	10	11	15	11	1	37	25	26	14	10	28	40	28	.	33
27	SK Vehicle	2.75	1.07	9	12	16	-9	4	8	1	8	2	-15	-3	-3	7	9	9	24	2	25	27	27	11	6	40	52	36	33	.

N= 401

Table 4

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 19E: ARMOR CREWMAN

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	Overall Rating	4.62	0.78	.	84	72	58	72	53	69	61	12	20	8	-37	41	12	15	16	4	25	23	22	19	10	27	26	18	22	12	7
2	Eff/Ldr Rating	4.38	0.74	84	.	68	55	76	50	80	65	16	21	8	-32	41	17	26	19	17	31	34	31	27	14	33	32	23	22	23	11
3	Discipline Rtng	4.50	0.83	72	68	.	45	53	41	55	64	-1	12	-14	-35	38	6	15	14	2	21	23	18	17	6	27	18	8	22	14	-2
4	Fitness Rating	4.76	0.82	58	55	45	.	44	39	43	36	10	43	-0	-19	28	8	2	16	-3	1	5	10	5	-4	0	-0	-0	4	2	2
5	Job-Spec Tech	23.19	3.20	72	76	53	44	.	75	71	55	10	14	17	-31	34	23	17	19	13	25	22	26	19	8	27	25	18	15	20	9
6	Job-Spec Other	14.71	1.89	53	50	41	39	75	.	50	41	9	13	13	-18	19	15	9	13	2	8	7	12	2	4	15	12	9	9	9	1
7	Combat Exmplry	8.38	1.36	69	80	55	43	71	50	.	63	15	18	8	-32	34	15	27	15	14	20	23	19	20	7	22	25	19	10	13	2
8	Combat Problems	9.80	1.47	61	65	64	36	55	41	63	.	-1	7	4	-31	29	13	22	13	6	24	18	21	13	8	24	18	17	15	16	-1
9	Awards & Certs	2.52	1.60	12	16	-1	10	10	9	15	-1	.	15	19	-7	13	6	4	-3	13	5	7	-0	10	-2	12	12	3	4	3	3
10	Phys. Readiness	249.41	27.11	20	21	12	43	14	13	18	7	15	.	-1	-10	10	-3	-3	4	2	-6	0	-6	4	1	-4	1	2	-2	2	-7
11	M16 Qualific.	2.40	0.68	8	8	-14	-0	17	13	8	4	19	-1	.	14	-1	7	7	3	10	11	12	13	17	31	10	6	12	2	16	-1
12	Articles 15	0.35	0.77	-37	-32	-35	-19	-31	-18	-32	-31	-7	-10	14	.	-43	-9	-8	-16	1	-13	-17	-17	-7	1	-19	-13	-12	-0	-7	-4
13	Promotion Rate	0.03	0.58	41	41	38	28	34	19	34	29	13	10	-1	-43	.	10	7	15	12	14	24	28	21	2	17	22	18	6	15	1
14	MO Tech.	50.00	9.99	12	17	6	8	22	15	15	13	6	-3	7	-9	10	.	18	24	20	36	27	27	13	18	23	18	9	2	19	0
15	MO Basic	38.16	2.48	15	26	15	2	17	9	27	22	4	-3	7	-8	7	18	.	21	23	30	32	25	21	18	21	25	11	4	19	-5
16	MO Safety	21.85	2.95	16	19	14	16	19	13	15	13	-3	6	3	-16	15	24	21	.	14	22	18	18	10	6	15	13	5	5	17	3
17	MO Comm	28.55	7.59	4	17	2	-3	13	2	14	6	13	2	10	1	12	20	23	14	.	23	28	25	32	11	20	23	13	3	22	2
18	AK Tech.	50.00	9.99	25	31	21	1	25	8	20	24	5	-6	11	-13	14	36	30	22	23	.	60	52	45	34	64	60	44	38	42	7
19	AK Basic	42.16	7.28	23	34	23	5	23	7	23	18	7	0	12	-17	24	27	32	18	28	60	.	65	53	38	65	67	46	41	43	6
20	AK Safety	21.19	4.13	22	31	18	10	26	12	19	21	-0	-6	13	-17	28	27	25	18	25	52	65	.	44	34	46	51	37	26	22	5
21	AK Comm	11.33	3.59	19	27	17	5	19	2	20	13	10	4	17	-7	21	13	21	10	22	45	53	44	.	16	45	51	34	30	24	2
22	AK Identify	10.05	1.78	10	14	6	-4	8	4	7	8	-2	1	31	1	2	18	18	6	11	34	30	34	16	.	24	28	22	18	27	3
23	SK Tech.	54.54	9.66	27	33	27	0	27	15	22	26	12	-4	10	-19	17	23	21	15	20	64	65	46	45	24	.	78	53	59	18	21
24	SK Basic	34.94	8.44	26	32	18	-0	25	12	25	18	12	1	6	-13	22	18	25	13	23	60	67	51	51	23	78	.	68	47	37	12
25	SK Safety	8.18	2.14	18	23	8	-0	18	8	19	17	3	2	12	-13	18	9	11	5	13	44	46	37	34	22	53	58	.	38	22	4
26	SK Comm	7.59	1.80	22	22	22	4	15	9	10	15	4	-2	2	-0	6	2	4	5	3	38	41	26	30	18	59	47	38	.	24	14
27	SK Vehicle	0.54	0.50	7	11	-2	2	6	1	2	-1	8	-7	-1	-6	1	0	-0	6	3	7	6	5	2	3	21	12	4	14	9	.
28	SK Identify	3.01	0.96	18	23	14	2	20	9	18	16	8	2	16	-7	15	19	19	17	23	42	43	33	24	37	48	47	33	24	.	9

N= 335

Table 5

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 31C: SINGLE CHANNEL RADIO OPERATOR

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	Overall Rating	4.73	0.79	.	83	73	64	74	66	66	66	17	11	2	-31	30	20	24	15	15	-2	24	17	9	14	3	13	19	10	14	2	-2
2	Eff/Ldr Rating	4.48	0.72	83	.	68	57	81	71	68	63	18	12	7	-31	30	24	21	21	15	2	30	28	14	16	6	13	23	12	20	12	4
3	Discipline Rtng	4.64	0.88	73	68	.	52	54	58	53	60	4	4	-11	-32	26	10	14	7	10	-1	20	19	4	15	6	7	9	-4	10	-2	-8
4	Fitness Rating	5.05	0.88	64	57	52	.	47	40	42	42	11	34	-6	-25	24	12	8	10	4	-2	1	2	0	8	-4	6	-5	-6	-2	-9	-12
5	Job-Spec Tech	14.27	2.01	74	81	54	47	.	76	66	57	14	4	5	-16	22	20	24	20	11	-1	29	30	16	15	8	15	19	15	16	13	-1
6	Job-Spec Other	14.37	2.09	66	71	58	40	76	.	54	48	3	-3	-3	-17	22	11	18	15	1	0	17	22	8	9	2	5	5	2	9	3	-6
7	Combat Exmplry	9.09	1.54	66	68	53	42	66	54	.	77	11	1	5	-21	17	6	13	18	23	-7	26	30	15	19	11	12	18	9	14	10	7
8	Combat Problems	10.47	1.71	66	63	60	42	57	48	77	.	9	-1	-2	-22	14	4	16	11	15	-0	22	24	3	14	-1	5	15	5	9	0	-3
9	Awards & Certs	2.16	1.75	17	18	4	11	14	3	11	9	.	23	10	2	12	9	12	6	3	2	10	10	11	-0	-5	8	8	12	4	4	6
10	Phys. Readiness	259.54	29.59	11	12	4	34	4	-3	1	-1	23	.	4	-11	6	1	-10	0	1	-6	-4	-8	4	1	3	-8	-4	-0	1	-13	-5
11	M16 Qualific.	2.16	0.77	2	7	-11	-6	5	-3	5	-2	10	4	.	4	3	4	5	10	7	5	7	10	8	-4	-6	5	9	4	4	11	10
12	Articles 15	0.34	0.84	-31	-31	-32	-25	-16	-17	-21	-22	2	-11	4	.	-34	-9	-3	-7	-12	-3	-16	-9	-13	-20	-10	-3	-11	-4	-12	-4	-3
13	Promotion Rate	-0.02	0.56	30	30	26	24	22	22	17	14	12	4	3	-34	.	9	12	21	9	5	18	17	10	19	13	12	13	15	12	4	-9
14	MO Tech.	78.44	9.69	20	24	10	12	20	11	6	4	9	1	4	-9	8	.	25	25	28	9	42	21	23	21	22	15	39	21	24	9	8
15	MO Basic	21.25	3.84	24	21	14	8	24	18	13	16	12	-10	5	-3	12	25	.	18	27	8	31	31	18	15	5	21	27	24	27	10	15
16	MO Safety	20.15	3.99	15	21	7	10	20	15	18	11	6	0	10	-7	21	25	18	.	23	16	10	21	13	9	6	8	11	10	19	4	9
17	MO Comm	16.73	6.39	15	15	10	4	11	1	23	15	3	1	7	-12	9	28	27	23	.	1	34	29	21	28	21	23	26	17	11	5	25
18	MO Vehicle	11.73	1.31	-2	2	-1	-2	-1	0	-7	-0	2	-6	5	-3	5	9	8	16	1	.	11	9	10	2	-6	7	22	16	14	12	11
19	JK Tech.	57.16	11.68	24	30	20	1	29	17	26	22	10	-4	7	-16	18	42	31	10	34	11	.	60	59	60	37	33	72	49	50	44	28
20	JK Basic	22.12	4.81	17	28	15	2	30	22	30	24	10	-8	10	-9	17	21	31	21	29	9	60	.	58	50	22	31	49	42	43	40	20
21	JK Safety	23.31	4.63	9	14	4	0	16	8	15	3	11	4	8	-13	10	23	18	13	21	10	59	58	.	50	28	30	44	40	48	26	24
22	JK Comm	10.12	2.74	14	16	15	8	15	9	19	14	-0	1	-4	-20	19	21	15	9	38	2	60	50	50	.	32	19	44	36	36	27	19
23	JK Vehicle	4.54	1.82	3	6	6	-4	8	2	11	-1	-5	3	-6	-10	13	22	5	6	21	-6	37	23	28	32	.	17	20	14	16	13	11
24	JK Identify	6.72	2.13	13	13	7	6	15	5	12	5	8	-6	5	-3	12	15	21	8	23	7	33	31	30	19	17	.	27	28	21	11	40
25	SK Tech.	77.87	15.43	19	23	9	-5	19	5	18	15	8	-4	9	-11	13	39	27	11	26	22	72	49	44	44	30	27	.	62	55	46	29
26	SK Basic	10.95	2.74	10	12	-4	-8	15	2	9	5	12	-0	4	-4	15	21	24	10	17	16	49	42	40	36	14	28	42	.	56	42	26
27	SK Safety	11.08	2.81	14	20	10	-2	16	9	14	9	4	1	4	-12	12	24	27	19	11	14	50	43	48	36	16	21	58	56	.	41	27
28	SK Vehicle	3.33	1.84	2	12	-2	-9	13	3	10	0	-4	-13	11	-4	4	9	10	4	5	12	44	40	36	27	13	11	48	42	41	.	22
29	SK Identify	1.16	0.93	-2	4	-8	-12	-1	-9	7	-3	6	-5	10	-3	-0	8	15	9	28	11	28	20	24	19	11	40	29	25	27	22	.

N= 239

Table 6

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 63B: LIGHT WEIGHT VEHICLE MECHANIC

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	Overall Rating	4.55	0.84	.	86	75	57	75	75	68	65	20	7	-4	-24	24	11	-1	5	10	20	15	22	11	21	22	21	15	19
2	Eff/Ldr Rating	4.31	0.83	86	.	75	50	84	78	69	66	21	1	-5	-23	23	19	-1	3	12	23	16	27	18	26	22	19	14	22
3	Discipline Rtng	4.54	0.88	75	75	.	51	63	65	59	66	19	2	-8	-27	26	10	-5	7	9	11	5	19	3	14	23	20	13	14
4	Fitness Rating	4.82	0.86	57	50	51	.	38	49	44	41	13	31	2	-20	20	-2	-2	8	7	-0	2	8	-0	-2	16	14	12	6
5	Job-Spec Tech	22.42	4.10	75	84	63	38	.	78	65	57	21	-1	-5	-16	16	23	1	3	13	28	21	26	19	37	21	16	6	28
6	Job-Spec Other	23.19	3.52	75	78	65	49	78	.	68	55	18	5	-8	-18	17	12	4	4	12	18	17	22	13	21	16	18	9	20
7	Combat Exmplry	8.87	1.61	68	69	59	44	65	68	.	69	14	4	-7	-16	17	13	0	9	9	16	11	23	8	29	18	14	6	13
8	Combat Problems	9.92	1.86	65	66	66	41	57	55	69	.	14	-0	-6	-20	27	10	-3	4	9	17	11	20	7	19	21	18	18	18
9	Awards & Certs	2.31	1.81	20	21	15	13	21	18	14	14	.	4	2	-11	7	11	-5	-0	7	7	2	12	11	13	14	10	8	15
10	Phys. Readiness	255.47	31.93	7	1	2	31	-1	5	4	-0	4	.	10	-10	15	1	8	3	-1	-7	-12	-2	-9	-10	1	0	-2	-4
11	M16 Qualific.	2.19	0.73	-4	-5	-8	2	-5	-8	-7	-6	2	10	.	1	-9	-2	5	-4	-0	-6	3	3	2	-2	-2	2	-0	4
12	Articles 15	0.37	0.85	-24	-23	-27	-20	-16	-18	-16	-20	-11	-10	1	.	-36	-3	-2	-2	-4	-7	-5	-6	-0	-6	-11	-7	-13	-8
13	Promotion Rate	0.04	0.52	24	23	26	20	16	17	17	27	7	15	-9	-36	.	-5	-4	-2	-1	13	9	4	8	13	16	15	9	12
14	MD Tech.	110.11	6.84	11	19	10	-2	23	12	13	10	11	1	-2	-3	-5	.	8	6	18	33	23	19	22	37	19	16	4	24
15	MD Basic	34.96	4.09	-1	-1	-5	-2	1	4	0	-3	-5	8	5	-2	-4	8	.	10	7	6	12	14	12	10	7	15	-1	14
16	MD Safety	21.92	3.25	5	3	7	8	3	4	9	4	-0	3	-4	-2	-2	6	10	.	2	2	5	18	1	1	2	7	-7	-9
17	MD Vehicle	11.22	1.84	10	12	9	7	13	12	9	9	7	-1	-0	-4	-1	18	7	2	.	15	6	4	11	17	6	6	2	13
18	JK Tech.	66.61	11.93	20	23	11	-0	28	18	16	17	7	-7	-6	-7	13	33	6	2	15	.	62	47	62	67	50	39	36	59
19	JK Basic	24.36	4.69	15	16	5	2	21	17	11	11	2	-12	3	-5	9	23	12	5	6	62	.	45	44	47	41	36	22	44
20	JK Safety	18.91	3.05	22	27	19	8	26	22	23	20	12	-2	3	-6	4	19	14	18	4	47	45	.	35	40	36	33	20	39
21	JK Vehicle	15.81	4.03	11	18	3	-0	19	13	8	7	11	-9	2	-0	8	22	12	1	11	62	44	38	.	56	37	21	24	48
22	SK Tech.	56.00	12.89	21	26	14	-2	37	21	20	19	13	-10	-2	-6	13	37	10	1	17	67	47	40	56	.	52	47	30	69
23	SK Basic	16.56	4.24	22	28	23	16	21	18	18	21	14	1	-2	-11	16	19	7	2	6	50	41	36	37	52	.	51	30	56
24	SK Safety	6.02	1.74	21	19	20	14	18	18	14	18	10	0	2	-7	15	16	15	7	6	39	36	33	21	47	51	.	39	50
25	SK Comm	0.90	0.30	15	14	12	13	6	9	8	18	8	-3	-0	-13	8	4	-1	-7	2	36	22	20	24	30	30	39	.	38
26	SK Vehicle	24.10	5.54	19	22	14	8	28	20	13	18	18	-4	4	-8	13	24	14	-0	13	59	44	39	49	69	56	50	39	.

N= 403

Table 7

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 64C: MOTOR TRANSPORT OPERATOR

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	Overall Rating	4.52	0.76	.	86	78	63	72	59	68	58	13	11	4	-30	33	8	20	15	16	17	19	3	13	12	16	14
2	Eff/Ldr Rating	4.36	0.75	86	.	77	59	78	69	74	56	17	9	6	-25	31	16	20	18	23	22	26	7	21	17	15	21
3	Discipline Rtng	4.53	0.81	78	77	.	52	67	51	58	54	10	3	-2	-29	35	4	14	14	15	15	19	1	16	12	15	16
4	Fitness Rating	4.74	0.87	63	59	52	.	54	39	46	35	3	28	3	-20	21	-2	8	14	5	6	4	2	5	6	7	-2
5	Job-Spec Tech	29.61	3.76	72	78	67	54	.	78	65	52	13	6	7	-21	25	9	16	19	17	20	19	5	16	17	12	15
6	Job-Spec Other	17.79	2.52	59	69	51	39	78	.	63	41	18	4	13	-15	19	12	11	16	17	16	19	4	13	17	7	14
7	Combat Exmplry	8.80	1.45	68	74	58	46	65	63	.	65	12	6	11	-21	22	20	19	16	20	15	20	5	18	8	10	15
8	Combat Problems	9.50	1.63	58	58	54	35	52	41	65	.	8	-3	2	-24	26	12	15	10	16	17	22	7	15	14	20	19
9	Awards & Certs	3.12	2.08	13	17	10	3	13	18	12	8	.	6	11	5	12	8	4	5	-3	-2	1	6	3	4	-1	2
10	Phys. Readiness	248.48	37.70	11	9	3	28	6	4	6	-3	6	.	3	-6	-1	-1	3	2	-4	-4	-4	2	-2	-5	0	-8
11	M16 Qualific.	2.09	0.75	4	6	-2	3	7	13	11	2	11	3	.	4	-5	9	12	5	7	5	3	-0	-1	-3	1	-1
12	Articles 15	0.49	0.98	-30	-25	-29	-20	-21	-15	-21	-24	5	-6	4	.	-36	-1	-11	-11	-7	-13	-12	0	-5	-8	-12	-4
13	Promotion Rate	-0.01	0.57	33	31	35	21	25	19	22	26	12	-1	-5	-36	.	10	9	10	9	12	11	5	11	9	8	11
14	MO Basic	43.44	10.16	8	16	4	-2	9	12	20	12	8	-1	9	-1	10	.	29	10	44	31	30	7	28	21	6	22
15	MO Safety	93.73	9.84	20	20	14	8	16	11	19	15	4	3	13	-11	9	29	.	14	27	31	24	4	24	19	14	24
16	MO Vehicle	33.30	4.19	15	18	14	14	19	16	16	10	5	2	5	-11	10	10	14	.	5	8	15	3	10	11	1	11
17	JK Basic	27.28	5.82	16	23	15	5	17	17	20	16	-3	-4	7	-7	9	44	27	5	.	67	54	10	47	39	20	49
18	JK Safety	33.42	5.42	17	22	15	6	20	16	15	17	-2	-4	5	-13	12	31	31	8	67	.	49	4	42	47	23	49
19	JK Vehicle	35.40	7.70	19	26	19	4	19	19	20	22	1	-4	3	-12	11	30	24	15	54	49	.	11	49	40	27	55
20	JK Identify	2.15	1.41	3	7	1	2	5	4	5	7	6	2	-0	0	5	7	4	3	10	4	11	.	17	10	-2	12
21	SK Basic	16.41	4.36	13	21	16	5	16	13	18	15	3	-2	-1	-5	11	28	24	10	47	42	49	17	.	56	43	38
22	SK Safety	6.44	1.93	12	17	12	6	17	17	8	14	4	-5	-3	-8	9	21	19	11	39	47	40	10	56	.	36	59
23	SK Comm	0.89	0.32	16	15	15	7	12	7	10	20	-1	0	1	-12	8	6	14	1	20	23	27	-2	43	36	.	37
24	SK Vehicle	55.72	10.07	14	21	16	-2	15	14	15	19	2	-9	-1	-4	11	32	24	11	49	49	55	13	58	59	37	.

N= 477

Table 8

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 71L: ADMINISTRATIVE CLERK

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	Overall Rating	4.92	0.85	.	83	71	57	72	63	63	59	20	24	4	-23	20	17	14	3	22	15	17	21	13	11	5	10
2	Eff/Ldr Rating	4.64	0.78	83	.	73	56	73	65	70	60	21	19	2	-19	19	25	14	2	29	17	18	29	17	9	7	11
3	Discipline Rtng	5.01	0.88	71	73	.	47	63	55	58	58	13	13	4	-27	19	20	10	-3	22	15	11	20	7	8	1	4
4	Fitness Rating	5.23	0.89	57	56	47	.	40	39	55	49	20	35	5	-23	20	3	7	-3	1	2	2	-1	0	-5	0	-2
5	Job-Spec Tech	19.88	2.73	72	73	63	40	.	76	54	50	8	7	-5	-21	21	24	8	-2	28	16	16	28	10	9	6	7
6	Job-Spec Other	18.57	3.13	63	65	55	39	76	.	50	46	10	13	-1	-21	17	22	13	1	22	15	16	26	8	9	10	8
7	Combat Exmplry	8.74	1.83	63	70	58	55	54	50	.	72	24	19	8	-15	18	9	20	11	13	23	17	14	13	8	8	23
8	Combat Problems	10.72	1.95	59	60	58	49	50	46	72	.	21	16	7	-22	13	12	14	6	11	26	12	15	13	9	1	14
9	Awards & Certs	2.62	1.73	20	21	13	20	8	10	24	21	.	17	20	-4	9	-0	10	-1	-0	5	11	-0	-2	-2	5	1
10	Phys. Readiness	260.40	33.39	24	19	13	35	7	13	19	16	17	.	11	-9	5	1	6	5	0	-5	8	5	4	12	2	8
11	M16 Qualific.	1.86	0.80	4	2	4	5	-5	-1	8	7	20	11	.	3	2	-4	12	8	-6	7	3	-3	2	-7	-1	13
12	Articles 15	0.22	0.62	-23	-19	-27	-23	-21	-21	-15	-22	-4	-9	3	.	-42	-13	-5	1	-10	-7	2	-10	-5	-5	-5	4
13	Promotion Rate	0.01	0.46	20	19	19	20	21	17	18	13	9	5	2	-42	.	12	5	2	6	6	9	9	7	6	4	-0
14	MO Tech.	96.09	14.26	17	25	20	3	24	22	9	12	-0	1	-4	-13	12	.	28	13	58	34	33	58	25	23	7	11
15	MO Basic	18.56	5.00	14	14	10	7	8	13	20	14	10	6	12	-5	5	28	.	43	29	48	35	23	26	17	6	23
16	MO Safety	20.54	4.00	3	2	-3	-3	-2	1	11	6	-1	5	8	1	2	13	43	.	11	28	23	7	13	10	0	17
17	JK Tech.	42.21	9.53	22	29	22	1	28	22	13	11	-0	0	-6	-10	6	58	29	11	.	47	48	73	42	24	17	17
18	JK Basic	25.23	5.16	15	17	15	2	16	15	23	26	5	-5	7	-7	6	34	48	28	47	.	50	40	44	27	27	28
19	JK Safety	16.24	3.01	17	18	11	2	16	16	17	12	11	8	3	2	9	33	35	23	48	50	.	43	38	32	19	25
20	SK Tech.	44.99	9.78	21	28	20	-1	28	26	14	15	-0	5	-3	-10	5	58	23	7	73	40	43	.	44	33	15	16
21	SK Basic	9.90	2.28	13	17	7	0	10	8	13	13	-2	4	2	-5	7	25	26	13	42	44	38	44	.	32	18	31
22	SK Safety	4.26	1.29	11	9	8	-5	9	9	8	9	-2	12	-7	-5	6	23	17	10	24	27	32	33	32	.	4	15
23	SK Comm	0.38	0.48	5	7	1	0	6	10	8	1	5	2	-1	-5	4	7	6	0	17	27	19	15	18	4	.	11
24	SK Vehicle	2.71	1.21	10	11	4	-2	7	9	23	14	1	8	13	4	-0	11	22	17	17	28	25	16	31	15	11	.

N= 353

Table 9

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 91A: MEDICAL SPECIALIST

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	Overall Rating	4.61	0.82	.	86	78	60	67	62	71	70	22	15	-2	-29	32	17	6	13	28	24	22	4	8	28	12	15	5
2	Eff/Ldr Rating	4.40	0.77	86	.	76	56	73	67	73	71	24	13	-4	-30	33	20	9	19	26	25	21	-2	13	23	14	16	9
3	Discipline Rtng	4.54	0.91	78	76	.	47	60	47	58	69	12	7	-8	-29	31	15	11	13	28	21	20	-4	6	33	14	11	10
4	Fitness Rating	4.74	0.92	60	56	47	.	41	38	49	47	10	39	0	-20	18	3	-0	-0	3	7	4	7	1	-1	2	-4	-15
5	Job-Spec Tech	23.09	3.24	67	73	60	41	.	67	55	54	15	6	-1	-27	26	18	2	13	22	16	14	-3	3	22	5	15	7
6	Job-Spec Other	18.47	2.55	62	67	47	38	67	.	64	51	28	7	9	-17	27	10	6	16	18	25	20	5	15	23	11	16	16
7	Combat Exmplry	9.20	1.48	71	73	58	49	55	64	.	79	30	9	9	-20	26	16	10	15	22	25	22	1	18	28	20	17	12
8	Combat Problems	10.11	1.77	70	71	69	47	54	51	79	.	23	5	-5	-28	30	14	6	11	24	22	23	-1	9	22	28	16	12
9	Awards & Certs	3.04	2.01	22	24	12	10	15	28	30	23	.	14	34	-6	13	3	7	22	4	10	8	11	16	4	11	12	6
10	Phys. Readiness	255.71	31.94	15	13	7	39	6	7	9	5	14	.	17	-11	-2	4	-6	-5	-3	-7	-2	3	-5	-6	-3	-6	-7
11	M16 Qualific.	2.08	0.78	-2	-4	-8	0	-1	9	9	-5	34	17	.	-1	-4	3	0	8	-8	5	-7	-0	12	-4	-2	2	2
12	Articles 15	0.41	0.89	-29	-30	-29	-20	-27	-17	-20	-28	-8	-11	-1	.	-33	-10	1	-7	-10	-7	-6	12	-5	-13	-15	-6	-1
13	Promotion Rate	-0.00	0.58	32	33	31	18	26	27	26	30	13	-2	-4	-33	.	10	9	7	16	20	9	-9	11	18	11	14	11
14	MD Tech.	50.48	10.02	17	20	15	3	18	10	16	14	3	4	3	-10	10	.	16	34	39	27	30	2	13	44	8	28	14
15	MD Basic	9.57	3.00	6	9	11	-0	2	6	10	6	7	-6	0	1	9	16	.	17	21	37	21	9	14	17	18	22	11
16	MD Safety	33.52	4.30	13	19	13	-0	13	16	15	11	22	-5	8	-7	7	34	17	.	32	35	33	3	17	30	10	23	13
17	MK Tech.	85.32	13.71	28	26	28	3	22	18	22	24	4	-3	-8	-10	16	39	21	32	.	54	78	13	16	67	20	46	22
18	MK Basic	15.19	3.63	24	25	21	7	16	25	25	22	10	-7	5	-7	20	27	37	32	54	.	55	8	24	41	23	38	22
19	MK Safety	42.71	7.25	25	21	20	4	14	20	22	23	8	-2	-7	-6	9	30	21	33	78	55	.	12	16	55	21	49	21
20	MK Vehicle	2.42	1.04	4	-2	-4	7	-3	5	1	-1	11	3	-0	12	-9	2	9	3	13	8	12	.	10	2	-3	6	6
21	MK Identify	6.62	2.32	8	13	6	1	3	15	18	9	18	-5	12	-5	11	13	14	17	16	24	16	10	.	15	15	13	13
22	SK Tech.	91.65	17.57	28	33	33	-1	32	23	29	32	4	-8	-4	-13	18	44	17	30	67	41	55	2	15	.	24	52	36
23	SK Basic	2.04	0.78	12	14	14	2	5	11	20	28	11	-3	-2	-16	11	8	18	10	20	23	21	-3	15	24	.	26	14
24	SK Safety	5.77	1.56	15	16	11	-4	15	18	17	16	12	-8	2	-6	14	28	22	33	48	38	49	6	12	52	26	.	27
25	SK Vehicle	4.51	1.62	6	9	10	-15	7	16	12	12	6	-7	2	-1	11	14	11	18	22	22	21	6	13	36	14	27	.

N= 372

Table 10

JOB PERFORMANCE MEASURE SUMMARY STATISTICS

FOR 95B: MILITARY POLICE

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	Overall Rating	4.74	0.80	.	87	69	70	78	68	74	70	18	22	13	-28	21	15	18	8	4	1	12	10	8	4	9	8	19	8	7	6	8	-6
2	Eff/Ldr Rating	4.50	0.73	87	.	71	61	77	72	76	68	20	17	11	-22	19	14	21	10	10	1	10	13	7	7	15	10	18	13	12	6	9	-5
3	Discipline Rtng	4.71	0.77	69	71	.	46	65	48	55	63	6	7	4	-27	26	6	9	5	7	-3	11	10	6	8	12	12	15	18	14	6	10	-3
4	Fitness Rating	4.90	0.84	70	61	46	.	58	56	55	52	16	43	13	-26	16	9	12	7	5	2	0	3	-2	-0	3	-3	3	5	3	-1	1	-7
5	Job-Spec Tech	29.00	3.66	78	77	65	58	.	73	68	63	15	15	11	-19	17	12	19	8	12	2	14	16	9	5	9	7	17	11	12	5	5	-3
6	Job-Spec Other	23.60	3.10	68	72	48	56	73	.	71	61	32	18	27	-16	8	11	22	14	19	7	2	13	6	4	16	7	9	12	6	5	4	-2
7	Combat Exmplry	9.56	1.36	74	78	55	55	68	71	.	79	19	19	16	-17	14	17	19	14	14	6	10	17	11	7	15	6	19	15	9	9	8	1
8	Combat Problems	10.45	1.53	70	68	63	52	63	61	79	.	15	15	15	-28	21	14	16	11	10	-0	16	18	17	11	14	10	19	15	9	6	12	0
9	Awards & Certs	3.17	2.09	18	20	6	16	15	32	19	15	.	20	26	-3	11	8	16	6	8	11	-11	2	-1	7	-0	9	6	4	10	4	9	-0
10	Phys. Readiness	251.75	32.78	22	17	7	43	15	18	19	15	20	.	13	-12	7	-1	6	4	2	4	-6	1	-3	-3	-2	-12	-2	-8	-4	-3	2	-5
11	M16 Qualific.	2.28	0.76	13	11	4	13	11	27	16	15	26	13	.	1	-3	4	6	5	4	6	-3	7	3	2	-9	-1	1	-0	-2	-2	-1	4
12	Articles 15	0.27	0.70	-28	-22	-27	-26	-19	-16	-17	-28	-3	-12	1	.	-39	-4	-0	-8	-3	2	-8	-8	-5	-2	0	0	-7	-6	-3	5	-7	6
13	Promotion Rate	0.01	0.47	21	19	26	16	17	8	14	21	11	7	-3	-39	.	4	4	6	1	-3	15	15	16	10	6	2	0	10	7	1	2	-1
14	MO Tech.	31.58	4.63	15	14	6	9	12	11	17	14	8	-1	4	-4	4	.	18	12	6	11	13	11	10	7	3	5	14	7	3	10	6	-9
15	MO Basic	50.04	10.28	18	21	9	12	19	22	19	16	16	6	6	-0	4	18	.	20	21	18	18	34	26	21	17	5	12	27	22	12	15	-7
16	MO Safety	31.76	5.16	8	10	5	7	8	14	14	11	6	4	5	-8	6	12	20	.	9	15	10	20	21	21	12	9	15	17	15	9	10	-6
17	MO Comm	10.57	2.17	4	10	7	5	12	19	14	10	8	2	4	-2	1	6	21	9	.	31	14	21	13	30	14	7	9	21	16	17	11	-10
18	MO Vehicle	10.56	-1.63	1	1	-3	2	2	7	6	-0	11	4	6	2	-3	11	18	15	31	.	1	4	8	19	16	2	11	12	9	13	10	-1
19	JK Tech.	38.44	5.90	12	10	11	0	14	2	10	16	-11	-6	-3	-8	15	13	16	10	14	1	.	60	53	35	19	15	40	33	28	24	19	1
20	JK Basic	50.11	9.99	10	13	10	3	16	13	17	19	2	1	7	-8	15	11	34	20	21	4	60	.	60	51	32	22	38	49	46	35	31	4
21	JK Safety	25.52	4.55	8	7	6	-2	9	6	11	17	-1	-3	3	-5	16	10	26	21	12	8	53	60	.	40	24	20	26	37	38	27	28	0
22	JK Comm	13.54	4.62	4	7	8	-0	5	4	7	11	7	-3	2	-2	10	7	21	21	30	19	35	51	40	.	26	18	22	33	31	36	24	-3
23	JK Vehicle	2.03	1.19	9	15	12	3	9	16	15	14	-0	-2	-9	0	6	3	17	12	14	16	18	32	24	26	.	15	18	23	22	20	17	-4
24	JK Identify	6.98	2.29	8	10	12	-3	7	7	6	10	9	-12	-1	0	2	5	5	9	7	2	15	22	20	18	15	.	21	20	21	17	12	-2
25	SK Tech.	40.20	7.04	19	18	15	3	17	9	19	19	4	-2	1	-7	0	14	12	15	9	11	40	38	36	22	18	21	.	49	49	38	37	2
26	SK Basic	17.85	3.86	8	13	18	5	11	12	15	15	4	-8	-0	-6	10	7	27	17	21	12	33	49	37	33	23	20	49	.	60	40	44	-1
27	SK Safety	14.45	3.35	7	12	14	2	12	6	9	9	10	-4	-2	-3	7	3	23	18	16	9	28	46	38	31	23	21	49	60	.	39	40	-3
28	SK Comm	3.12	1.23	6	8	6	-1	5	5	9	6	4	-3	-2	5	1	10	12	9	17	13	24	35	27	26	20	17	38	40	39	.	32	-0
29	SK Vehicle	6.02	1.90	8	9	10	1	5	4	8	13	9	2	-1	-7	2	6	15	10	11	10	19	31	28	24	17	12	37	44	40	32	.	-1
30	SK Identify	0.29	0.51	-6	-5	-3	-7	-2	-2	1	0	-0	-5	4	6	-1	-0	-7	-6	-10	-1	1	4	0	-3	-4	-2	2	-1	-6	-0	-1	.

N= 506

Table 11

FACTOR LOADINGS

SEPARATE MODEL FOR EACH JOB

	Military Occupational Specialty								
Construct/Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Core Technical									
HO Tech	--	.61	.47	.64	.51	.29	.77	.59	.32
JK Tech	--	.75	.78	.79	.74	.26	.78	.75	.32
SK Tech	--	.70	.79	.73	.82	.55	.229	.81	.43
MOS Tech Rtng	--	.45	.10	.22	.25	.25	.34	.10	.13
General Soldiering									
HO Soldier	.60	.51	.46	.64	.17	.50	.60	.42	.60
HO Safety	.26	.33	.32	.31	.12	.63	.37	.48	.47
HO Comm	.05	.06	.39	.56	--	--	--	--	.80
HO Vehicle	--	--	--	.22	.17	**	--	--	.31
JK Soldier	.76	.52	.74	.62	.45	.48	.87	.58	.46
JK Safety	.55	.37	.75	.38	.71	.51	.72	.58	.33
JK Comm	.30	.23	.66	.38	--	--	--	--	.29
JK Vehicle	--	.17	--	.10	.41	**	--	--	.35
JK Identify	.46	--	.20	.28	--	.12	--	.24	.21
SK Soldier	.73	.45	.67	.39	.78	.56	.45	.44	.42
SK Safety	.47	.32	.53	.62	.57	.47	.30	.64	.32
SK Comm	.42	.26	.42	--	.41	.35	.20	--	.20
SK Vehicle	.22	.24	.05	.30	.61	**	.22	.47	.28
SK Identify	.46	--	.46	.13	--	--	--	--	--
Effort/Leadership									
Eff/Ldr Rating	.76	.56	.85	.64	.68	.83	.66	.76	.70
MOS Tech Rtngs	.70	--	.63	.40	.41	.50	.25	.59	.52
MOS Other Rtng	.77	.41	.48	.43	.54	.62	.43	.61	.56
Comb Emplry	.80	.47	.68	.54	.57	.87	.63	.80	.77
Comb Problems	.48	.20	--	.39	.52	.53	.55	--	.56
Awards/Cert	.32	.23	.24	.19	.28	.25	.34	.34	.22
Overall Rating	.46	.39	.33	.17	.57	.42	.65	--	.41

Table 11

FACTOR LOADINGS

SEPARATE MODEL FOR EACH JOB

(continued)

	Military Occupational Specialty								
Construct/Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Discipline									
Discpln Rtng	.77	.58	.73	.45	.63	.85	.74	.58	.73
Comb Problems	.29	.16	.62	.03	.05	.19	--	.82	.33
Articles 15	-.63	-.61	-.55	-.62	-.65	-.47	-.69	-.46	-.60
Promotion Rate	.74	.61	.68	.79	.63	.57	.59	.54	.54
Overall Rating	.39	.20	.53	.54	.09	.42	.06	.75	.38
Fitness/Bearing									
Fitness Ratngs	.69	.23	.84	.48	.54	.42	.50	.60	.78
Phys Readiness	.11	.90	.49	.89	.70	.53	.76	.69	.69
Ratings Method									
AW Ratings	.60	.73	.47	.70	.66	.54	.65	.66	.66
MOS Ratings	.73	.73	.60	.69	.67	.49	.69	.54	.63
Comb Ratings	.47	.65	.55	.69	.57	.27	.55	.47	.40
Written Method									
JK Tech	--	.47	.28	.55	.59	.73	.44	.58	.57
JK Soldier	.41	.51	.33	.40	.61	.57	.11	.37	.59
JK Safety	.37	.52	.12	.63	.08	.49	.17	.76	.57
JK Comm	.34	.11	.07	.55	--	--	--	--	.52
JK Vehicle	--	--	--	.42	.62	**	--	.24	.21
JK Identify	-.15	.23	.50	.36	--	.05	--	.08	.23
SK Tech	--	.48	.48	.55	.46	.88	.42	.27	.50
SK Soldier	.50	.66	.54	.59	.15	.51	.54	--	.54
SK Safety	.53	.55	.42	.29	.34	.48	.44	.19	.60
SK Comm	.51	.47	.46	--	.16	.24	.05	--	.42
SK Vehicle	.49	.57	.24	.48	.55	**	.38	.05	.42
SK Identify	.21	--	.42	.44	--	--	--	--	--
M16 Qualification	.71	.71	.71	.71	.71	.71	.71	.71	.71

** Vehicle content was merged into the Core Technical factor for 64C.

Table 12

UNIQUENESS ESTIMATES

SEPARATE MODEL FOR EACH JOB

	Military Occupational Specialty								
Factor Score	11B	13B	19E	31C	63B	64C	71L	91A	95B
HO Tech	--	.52	.71	.48	.64	.74	.33	.57	.88
HO Soldier	.59	.66	.75	.52	.95	.74	.55	.76	.63
HO Safety	.92	.85	.75	.52	.95	.59	.79	.71	.77
HO Comm	.95	.95	.81	.62	--	--	--	--	.82
HO Vehicle	--	--	--	.03	.95	**	--	--	.90
JK Tech	--	.21	.30	.15	.12	.39	.17	.11	.53
JK Soldier	.10	.43	.22	.26	.29	.74	.31	.58	.43
JK Safety	.32	.53	.32	.31	.45	.49	.44	.15	.57
JK Comm	.56	.93	.32	.34	--	--	--	--	.64
JK Vehicle	--	--	--	.56	.32	**	--	.94	.82
JK Identify	.36	.89	.40	.51	--	.95	--	.92	.90
SK Tech	--	.27	.13	.09	.10	.14	.14	.15	.52
SK Soldier	.09	.37	.14	.48	.31	.42	.54	.74	.46
SK Safety	.46	.59	.43	.41	.50	.55	.72	.47	.55
SK Comm	.40	.72	.35	--	.65	.82	.78	--	.67
SK Vehicle	.73	.62	.69	.55	.18	**	.73	.76	.75
SK Identify	--	.45	.10	.22	.25	.25	.34	.10	.13
Overall Rating	.13	.13	.13	.13	.13	.13	.13	.13	.18
Eff/Ldr Rating	.11	.11	.11	.11	.11	.05	.11	.11	.05
Discpln Rating	.22	.22	.22	.22	.22	.05	.22	.22	.06
Fitness Rating	.38	.38	.38	.38	.38	.05	.38	.38	.05
MOS Tech Rtngs	.08	.11	.13	.14	.08	.37	.17	.12	.33
MOS Other Rtng	.10	.13	.17	.19	.12	.35	.20	.18	.27
Comb Exmplry	.02	.02	.02	.02	.02	.14	.02	.02	.08
Comb Problems	.13	.13	.13	.13	.13	.60	.13	.13	.40
Awards/Cert	.89	.94	.93	.95	.91	.94	.86	.85	.90
Phys Readiness	.95	.33	.67	.34	.50	.83	.46	.49	.49
Articles 15	.58	.59	.68	.60	.56	.76	.51	.75	.64
Promotion Rate	.45	.60	.53	.41	.57	.64	.62	.67	.70
M16	.50	.50	.50	.50	.50	.50	.50	.50	.50

** Vehicle content was merged into the Technical factor for 64C.

Table 13

ESTIMATED CONSTRUCT CORRELATIONS

SEPARATE MODEL FOR EACH JOB

		Military Occupational Specialty							
1st Construct	2nd Construct	11B	13B	19E	31C	63B	64C	71L	91A	95B
Core Technical	Gen Soldiering	--	.77	.83	.63	.58	.73	.48	.66	.70
	Effort/Lead	.67	.86	.51	.44	.50	.78	.44	.35	.46
	Discipline	.42	.13	.37	.26	.12	.69	.19	.43	.50
	Fitness	.25	.01	.03	.04	-.18	-.09	.10	-.05	-.09
	M16	.27	.00	.04	.11	.05	.05	-.09	-.17	-.10
General Soldiering	Effort/Lead	--	.89	.58	.57	.53	.44	.37	.43	.40
	Discipline	--	.29	.45	.30	.29	.29	.04	.37	.24
	Fitness	--	-.19	.05	-.05	-.03	-.14	.09	-.05	.00
	M16	--	-.06	.30	.30	.04	.11	.27	.02	.02
Effort/ Leadership	Discipline	.49	.67	.62	.55	.65	.51	.51	.59	.39
	Fitness	.57	.04	.38	-.11	.10	.23	.32	.21	.42
	M16	.38	-.13	.21	.24	-.02	.35	.22	.17	.28
Discipline	Fitness	.33	.05	.24	.24	.30	.30	.27	.19	.25
	M16	-.12	-.25	-.30	.09	-.28	-.11	.01	-.28	-.08
Fitness	M16	.52	.26	-.05	.02	.19	.22	.18	.27	.26

Table 14

GOODNESS-OF-FIT INDICES

SEPARATE MODEL FOR EACH JOB

	MOS	Root Mean Square Residual	Chi-Square	df	p
11B:	Infantryman	.061	326.2	227	.02
13B:	Cannon Crewman	.057	350.0	322	.14
19E:	Tank Crewman	.065	170.0	348	.999
31C:	Radio/Teletype Operator	.069	369.2	375	.58
63B:	Vehicle/Generator Mechanic	.060	332.1	296	.07
64C:	Motor Transport Operator	.058	280.1	247	.07
71L:	Administrative Clerk	.067	232.6	249	.77
91A:	Medical Specialist	.061	277.1	275	.45
95B:	Military Police	.052	470.0	374	.001

Table 15

FACTOR LOADINGS

SINGLE MODEL ACROSS ALL JOBS

	Military Occupational Specialty								
Construct/Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Core Technical									
HO Tech	--	.59	.43	.58	.46	.27	.71	.54	.29
JK Tech	--	.71	.79	.76	.57	.72	.70	.74	.37
SK Tech	--	.66	.70	.54	.73	.55	.68	.85	.42
MOS Tech Rtng	--	.21	.12	.16	.25	.01	.12	.05	-.02
General Soldiering									
HO Soldier	.52	.66	.44	.52	.16	.51	.57	.35	.58
HO Safety	.20	.44	.31	.36	.10	.49	.30	.50	.41
HO Comm	.06	.12	.37	.52	--	--	--	--	.43
HO Vehicle	--	--	--	.15	.21	**	--	--	.27
JK Soldier	.95	.50	.79	.64	.42	.69	.66	.69	.49
JK Safety	.69	.36	.75	.45	.53	.66	.57	.65	.42
JK Comm	.35	.25	.59	.51	--	--	--	--	.39
JK Vehicle	--	--	--	.28	.37	**	--	.07	.34
JK Identify	.43	.21	.34	.36	--	.12	--	.39	.18
SK Soldier	.81	.40	.67	.33	.70	.50	.42	.40	.38
SK Safety	.57	.34	.45	.40	.63	.43	.31	.62	.34
SK Comm	.51	.21	.31	--	.42	.29	.17	--	.23
SK Vehicle	.35	.22	.06	.17	.65	**	.32	.36	.21
Effort/Leadership									
Eff/Ldr Rating*	.76	.76	.76	.76	.76	.76	.76	.76	.76
MOS Tech Rtngs*	.59	.33	.54	.50	.45	.62	.43	.62	.62
MOS Other Rtng*	.77	.59	.33	.45	.59	.48	.47	.58	.58
Comb Emplry*	.72	.72	.72	.72	.72	.72	.72	.72	.72
Comb Problem*	.44	.44	.44	.44	.44	.44	.44	.44	.44
Awards/Cert*	.26	.26	.26	.26	.26	.26	.26	.26	.26
Overall Rating*	.48	.48	.48	.48	.48	.48	.48	.48	.48

Table 15

FACTOR LOADINGS

SINGLE MODEL ACROSS ALL JOBS

(continued)

	Military Occupational Specialty								
Construct/Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Discipline									
Discpln Rtng*	.69	.69	.69	.69	.69	.69	.69	.69	.69
Comb Problems*	.25	.25	.25	.25	.25	.25	.25	.25	.25
Articles 15*	-.48	-.48	-.48	-.48	-.48	-.48	-.48	-.48	-.48
Promotion Rate*	.52	.52	.52	.52	.52	.52	.52	.52	.52
Overall Rating*	.28	.28	.28	.28	.28	.28	.28	.28	.28
Fitness/Bearing									
Fitness Ratngs*	.82	.82	.82	.82	.82	.82	.82	.82	.82
Phys Readiness*	.37	.37	.37	.37	.37	.37	.37	.37	.37
Ratings Method									
AW Ratings*	.56	.56	.56	.56	.56	.56	.56	.56	.56
MOS Ratings*	.61	.61	.61	.61	.61	.61	.61	.61	.61
Comb Ratings*	.42	.42	.42	.42	.42	.42	.42	.42	.42
Written Method									
JK Tech	--	.49	.29	.54	.71	.30	.42	.49	.49
JK Soldier	-.16	.51	.29	.40	.53	.25	.28	.60	.60
JK Safety	-.07	.49	.07	.52	.26	.28	.35	.52	.52
JK Comm	.00	.11	.19	.38	--	--	--	.41	.41
JK Vehicle	--	--	--	.19	.62	**	--	.20	.20
JK Identify	-.05	.20	.12	.17	--	.10	--	.25	.25
SK Tech	--	.54	.65	.64	.49	.71	.45	.53	.53
SK Soldier	.44	.68	.58	.61	.25	.66	.50	.60	.60
SK Safety	.34	.51	.49	.57	.18	.56	.30	.59	.59
SK Comm	.51	.46	.60	--	.20	.36	.20	.50	.50
SK Vehicle	.38	.51	.17	.60	.45	**	.17	.46	.46

* These loadings were constrained to be equal across all MOS.
** Vehicle content was merged into the Core Technical factor for 64C.

Table 16

UNIQUENESS ESTIMATES

SINGLE MODEL ACROSS ALL JOBS

	Military Occupational Specialty								
Factor Score	11B	13B	19E	31C	63B	64C	71L	91A	95B
HO Tech	--	.62	.79	.62	.76	.91	.44	.68	.90
HO Soldier	.72	.58	.80	.70	.95	.73	.64	.87	.67
HO Safety	.95	.84	.90	.87	.95	.73	.90	.75	.81
HO Comm	.95	.95	.86	.71	--	--	--	--	.82
HO Vehicle	--	--	--	.95	.95	**	--	--	.93
JK Tech	--	.23	.28	.13	.15	.32	.28	.16	.60
JK Soldier	.10	.44	.28	.40	.48	.41	.44	.47	.40
JK Safety	.48	.56	.41	.49	.62	.44	.55	.26	.54
JK Comm	.85	.91	.57	.55	--	--	--	--	.67
JK Vehicle	--	--	--	.87	.44	**	--	.95	.85
JK Identify	.71	.90	.84	.81	--	.95	--	.64	.90
SK Tech	--	.25	.10	.24	.18	.17	.27	.19	.54
SK Soldier	.13	.37	.20	.52	.41	.31	.58	.83	.49
SK Safety	.54	.62	.54	.51	.55	.51	.80	.29	.54
SK Comm	.46	.75	.48	--	.77	.78	.92	--	.70
SK Vehicle	.75	.68	.95	.61	.31	**	.86	.86	.75
Overall Rating*	.18	.18	.18	.18	.18	.18	.18	.18	.18
Eff/Ldr Rating*	.09	.09	.09	.09	.09	.09	.09	.09	.09
Discpln Rating*	.17	.17	.17	.17	.17	.17	.17	.17	.17
Fitness Rating*	.05	.05	.05	.05	.05	.05	.05	.05	.05
MOS Tech Rtngs*	.18	.34	.22	.24	.18	.18	.18	.18	.25
MOS Other Rtng*	.05	.24	.46	.37	.05	.05	.05	.05	.27
Comb Exmplry*	.26	.26	.26	.26	.26	.26	.26	.26	.26
Comb Problems*	.29	.29	.29	.29	.29	.29	.29	.29	.29
Awards/Cert*	.93	.93	.93	.93	.93	.93	.93	.93	.93
Phys Readiness*	.83	.83	.83	.83	.83	.83	.83	.83	.83
Articles 15*	.77	.77	.77	.77	.77	.77	.77	.77	.77
Promotion Rate*	.70	.70	.70	.70	.70	.70	.70	.70	.70

* These loadings were constrained to be equal across all MOS.
** Vehicle content was merged into the Core Technical factor for 64C.

Table 17

MAPPING OF PERFORMANCE FACTORS ONTO LATENT PERFORMANCE CONSTRUCTS

	Content Constructs					Method Constructs		M16
Criterion Measure*	Core Technical Proficiency	General Soldiering Proficiency	Effort/ Leadership	Personal Discipline	Physical Fitness/ Military Bearing	Written Knowledge Tests	Rating Scales	M16 Qualification
AVB Effort			X				X	
AVB Discipline				X			X	
AVB Fitness					X		X	
AVB Overall			X	X			X	
MOS Technical			X				X	
MOS Other			X				X	
Cmbt Perform Well			X				X	
Cmbt Avoid Mistake			X	X			X	
Adm Awards/Certs			X					
Adm Phys Readiness					X			
Adm M16								X
Adm Articles 15				X				
Adm Promotion Rate				X				
HO Technical	X							
HO Communications		X						
HO Vehicles		X						
HO General Soldier		X						
HO ID Threat/Target		X						
HO Safety/Survival		X						
JK Technical	X					X		
JK Communications		X				X		
JK Vehicles		X				X		
JK General Soldier		X				X		
JK ID Threat/Target		X				X		
JK Safety/Survival		X				X		
SK Technical	X					X		
SK Communications		X				X		
SK Vehicles		X				X		
SK General Soldier		X				X		
SK ID Threat/Target		X				X		
SK Safety/Survival		X				X		

Note: Within each rating instrument, all of the factors were constrained to have an equal loading on the Rating Scales method construct. For example, the Perform Well and Avoid Mistakes factors from the Combat Performance Prediction Scale were constrained to have identical loadings on the Rating Scales method construct, but this loading did not have to be the same as the loading for the Army-Wide BARS factors, the MOS-Specific BARS factors, or the Common Task Scales factors.

*AVB = Army wide behaviorally anchored rating scales; HO = hands-on; JK = job knowledge; SK = school knowledge.

Table 18

Mean Intercorrelations among 12 Summary Criterion Measures
for the Batch A ARB

Criterion Summary Score	ARBAMCIP	ARBAMSOP	ARBAMELS	ARBAMSPD	ARBAMPFB	ARBPCII	ARBHOOIT	ARBESCIP	ARBESSOP	ARBESELS	ARBESSPD	ARBESPFB
ARBAMCIP: Core Tech Prof (raw)	1.000	0.531	0.280	0.190	0.032	0.243	0.737	0.000	0.380	0.465	0.225	0.040
ARBAMSOP: Gen Soldier Prof (raw)	0.531	1.000	0.260	0.163	0.041	0.206	0.722	0.306	0.091	0.451	0.192	0.047
ARBAMELS: Effort/Leadership (raw)	0.280	0.260	1.000	0.590	0.457	0.068	0.261	0.351	0.328	0.647	0.206	0.107
ARBAMSPD: Personal Discipline (raw)	0.190	0.163	0.590	1.000	0.335	0.650	0.150	0.256	0.226	0.436	0.804	0.194
ARBAMPFB: Fitness/Bearing (raw)	0.032	0.041	0.457	0.335	1.000	0.474	0.071	0.031	0.042	0.253	0.173	0.921
ARBPCII: Overall Perf Rating	0.243	0.206	0.068	0.650	0.474	1.000	0.203	0.309	0.261	0.444	0.333	0.192
ARBHOOIT: Hands-On Total	0.737	0.722	0.261	0.150	0.071	0.203	1.000	0.023	0.795	0.438	0.179	0.006
ARBESCIP: Core Tech Prof (resid)	0.000	0.306	0.351	0.256	0.031	0.309	0.023	1.000	0.440	0.453	0.246	-0.009
ARBESSOP: Gen Soldier Prof (resid)	0.380	0.091	0.328	0.226	0.042	0.261	0.795	0.440	1.000	0.432	0.212	0.007
ARBESELS: Effort/Leadership (resid)	0.465	0.451	0.647	0.436	0.253	0.444	0.438	0.453	0.432	1.000	0.477	0.277
ARBESSPD: Personal Discipline (resid)	0.225	0.192	0.206	0.804	0.173	0.333	0.179	0.246	0.212	0.477	1.000	0.200
ARBESPFB: Fitness/Bearing (resid)	0.040	0.047	0.107	0.194	0.921	0.192	0.006	-0.009	0.007	0.277	0.200	1.000

FIGURE 1

PRELIMINARY MODEL OF ENLISTED JOB PERFORMANCE

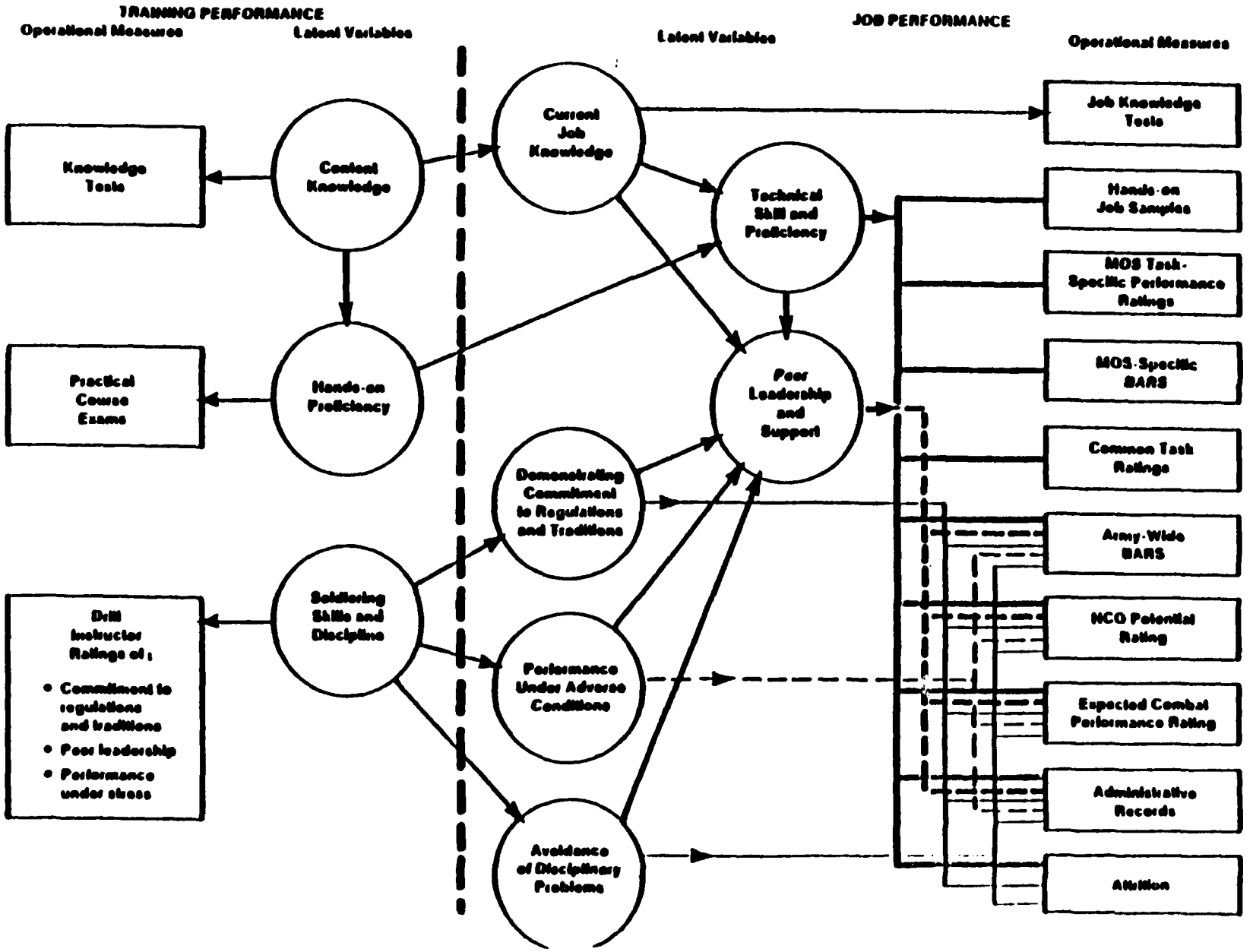


Figure 2

Functional Categories by Job and Method

Cluster Number and Name	11B			13B			19E			31C			63B			64C			71L			91A			95B		
	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK
1. First Aid	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2. Navigate		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3. NBC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X
4. Weapons	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X		X	X	X
5. Field Techniques	X	X	X	X	X	X		X	X		X	X		X	X		X	X		X	X		X	X	X	X	X
6. Communication	X	X	X	X	X	X	X	X	X	X	X				X			X			X				X	X	X
7. ID Target		X	X		X			X	X		X	X					X						X			X	X
8. Customs and Laws			X		X	X		X	X		X	X	X	X	X		X	X		X	X			X			X
9. Antitank/Antiair Weapons		X	X			X						X						X			X						X
11. Drive (Operate and Maintain)	X		X			X				X	X	X			X	X	X	X			X		X	X	X	X	X
14. Prepare/Operate/Maintain Howitzer and Ammunition				X	X	X																					
15. Operate Howitzer Sight/Alignment Device				X	X	X																					
16. Preventive Maintenance															X			X									X
17. Tank Operations							X	X	X																		
18. Tank Gunnery							X	X	X																		
20. Generators										X	X																
21. TTY Station and Net Operators										X	X																
22. Maintain TTY Electronic Equipment											X	X															
23. Operate TTY Electronic Equipment										X	X																
24. Install TTY Electronic Equipment										X	X																
25. Electrical System													X	X	X												
26. Brake/Steering/Suspension System													X	X	X												
27. Vehicle Operation and Recovery													X	X	X			X									X
28. Fuel/Cooling/Lubricating													X	X	X												
29. Forms/Files Management																			X	X	X						
30. Supervision/Coordination																				X	X						
31. Correspondence																			X	X	X						
32. Classified Material																			X	X	X						
33. Clinic/Ward Treatment and Care																						X	X	X			
34. Clinic/Ward Housekeeping																						X	X	X			
35. Clinic/Ward Management																							X	X			
36. General Medical Knowledge																								X			
37. Responding To Alarms																										X	X
38. Conduct MP Procedures																										X	X
39. Patrol Duties																									X	X	X
93. Power Train and Clutch													X	X													

Figure 3

SAMPLE FUNCTIONAL DUTY CLUSTER DEFINITIONS

First Aid

Consists of items whose primary purpose is to indicate knowledge about how to sustain life, prevent health complications caused by trauma or environmentally induced illness, including the practice of personal hygiene. Includes all related diagnostic, transportation, and treatment items except those items normally performed in a patient care facility. Includes items related to safety and safety hazards.

Navigate

Consists of items whose primary purpose is to indicate knowledge about how to plan or execute movement between points over unknown terrain either cross-country or using road networks, or identify the location of objects. Includes all means of determining direction, distances, and locations using maps of all types, overlays, compasses, terrain, celestial objects, and field expedients.

NBC

Consists of items whose primary purpose is to indicate knowledge about performance when nuclear, biological, or chemical contaminants and threats are present, planned, detected, or expected. Includes maintenance and operation of clothing, gear, and equipment whose primary purpose is to counter, protect, or detect NBC threats. Includes NBC markers. Does not include first aid treatment of contamination.

Weapons

Consists of items whose primary purpose is to indicate knowledge about maintenance, preparation, and firing of small arms. Small arms are defined as sized weapons, including automatic weapons, up to and including caliber .60 and shotguns. Includes ancillary sighting systems and techniques, stands and mounts, zeroing and techniques of fire. Excludes firing from aircraft and vehicles where the weapon is fired by electrical/hydraulic aiming/firing systems and sighting systems that are part of the aircraft/vehicle and not part of the weapon.

Figure 4

DEFINITIONS OF THE PERFORMANCE CONSTRUCTS

- Core Technical Proficiency
  This performance construct represents the proficiency with which the soldier performs the tasks that are "central" to the MOS. The tasks represent the core of the job and they are the primary definers of the MOS. For example, the first tour Armor Crewman starts and stops the tank engines; prepares the loader's station; loads and unloads the main gun; boresights the M60A3; engages targets with the main gun; and performs misfire procedures. This performance construct does not include the individual's willingness to perform the task or the degree to which the individual can coordinate efforts with others. It refers to how well the individual can execute the core technical tasks the job requires, given a willingness to do so.

- General Soldiering Proficiency
  In addition to the core technical content specific to an MOS, individuals in every MOS also are responsible for being able to perform a variety of general soldiering tasks -- for example, determines grid coordinates on military maps; puts on, wears and removes M17 series protective mask with hood; determines a magnetic azimuth using a compass; collects/reports information - SALUTE; and recognizes and identifies friendly and threat aircraft. Performance on this construct represents overall proficiency on these general soldiering tasks. Again, it refers to how well the individual can execute general soldering tasks, given a willingness to do so.

- Effort and Leadership
  This performance construct 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. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient? While appropriate knowledges and skills are necessary for successful performance, this construct is only meant to reflect the individual's willingness to do the job required and to be cooperative and supportive with other soldiers.

Figure 4

(continued)

- **Personal Discipline**
  This performance construct reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems. People who rank high on this construct show a commitment to high standards of personal conduct.

- **Physical Fitness and Military Bearing**
  This performance construct represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

PREDICTING SOLDIER PERFORMANCE:

ASSESSMENT OF TEMPERAMENT CONSTRUCTS AS

PREDICTORS OF JOB PERFORMANCE

Leaetta M. Hough
Steven D. Ashworth

Personnel Decisions Research Institute

Presented at the Annual Conference of the
Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

Predicting Soldier Performance: Assessment of Temperament Constructs as Predictors of Job Performance

Personnel Decisions (PDRI), along with Human Resources Research Organization (HumRRO) and the American Institutes for Research (AIR), has been involved in research with the Army Research Institute (ARI) to augment the prediction of job performance of enlisted Army personnel. It is a major project involving four research institutes, several million dollars, several years of effort, and thousands of soldiers. I'm going to talk about a small part of that project--the development and validation of temperament predictors, our strategy and results obtained.

The topics that I'm going to cover today are:

I. Literature Review
   - A. Strategy
   - B. Results

II. Development and Evaluation of Temperament Scales--"Assessment of Background and Life Experiences" (ABLE)
   - A. Target constructs
   - B. ABLE scale characteristics
   - C. ABLE factor structure
   - D. Definition of criterion composites
   - E. Zero-order validities of temperament scales
   - F. Contribution compared to other predictors

III. Evaluation of Response Validity Scales
   - A. "Non-Random Response" scale
   - B. "Self-Knowledge" scale
   - C. "Social Desirability" scale
   - D. "Poor Impression"

Literature Review

The project began in 1982, about five years ago, with a thorough literature review to identify potentially useful predictors of criteria important to the Army.

Our approach was construct-oriented for both predictors and criteria. Thus, we needed a classification strategy or taxonomy for both predictors and criteria.

Predictor and criterion taxonomies. Our classification system for criteria was: (1) educational, (2) training, (3) job involvement, (4) job proficiency, and (5) adjustment. Within each of these broad taxonomies we had subcategories such as supervisory/teacher ratings, GPA, etc. For the predictors, we started with the structure initially found by Tupes and Christal (1961) in the early 60s in their factor analysis of peer ratings. These factors were essentially replicated by Norman (1963) in his work with peer ratings of temperament. These five factors are what is being referred to today as the "Big Five." Following Hogan's thinking (at that time when we did our literature review) we separated "Affiliation" from the "Surgency" construct. Thus, our taxonomy consisted of the following: (1) Surgency, (2) Affiliation, (3) Adjustment, (4) Agreeableness, (5) Dependability, and (6) Intellectance.

Categorization of temperament scales. Once we had a taxonomy, our next step was to categorize the existing temperament scales into the classification scheme. From articles and manuals, we obtained hundreds of correlations between temperament scales. We categorized the temperament scales into the six categories, plus a seventh miscellaneous category, and then refined the classifications through an iterative process of classifying and reclassifying temperament scales to maximize the mean within-

category correlations and minimize the mean between-category correlations. The results of this bootstrapping process is shown in Table 1. The circles in the diagonal show the mean within-category correlations. As can be seen, they are in the .30s and .40s and are, in all cases, higher than the mean between-category correlations. Mean correlations in the .30s and .40s, however, suggest that the categories are not all that homogeneous. We could have increased the mean within-category correlations by putting more scales in the miscellaneous category; that, however, would have defeated our purpose of summarizing criterion-related validities according to constructs.

Meta analysis of criterion-related validities. Our next step was to summarize the criterion-related validities according to these predictor and criterion constructs. This next page of your handout, Table 2, shows the results. It is a meta analysis of the criterion-related validities of scales within each predictor construct for each criterion construct. As you can see, several constructs correlate with the various criteria. Note that there are three additional predictor constructs. These three, "Achievement," "Masculinity," and "Locus of Control," were all a part of the miscellaneous category. When we summarized the validities for the miscellaneous category, we found respectable validities there too, so we looked more closely at the scales included in the miscellaneous category and found these additional three constructs. We summarized the validities separately for these three constructs. Thus, in terms of criterion-related validities, the five basic constructs did not appear to cover the domain.

The results in this table are different from the results that Guion and Gottier published in their 1965 Personnel Psychology article; their conclusions were quite discouraging. They concluded that temperament vari-

ables have validity more often than can be expected by chance, but that no generalized principles can be discerned from the overall results. We believe that our strategy of summarizing the validities according to both predictor and criterion constructs accounts for the difference in results. The constructs provide the "generalized principles." To test this hypothesis, we summarized the validity coefficients in our database without regard to construct and obtained a coefficient of essentially zero, quite different from the coefficients in Table 2. We believe this demonstrates the importance of constructs as organizing principles for examining and understanding the literature on the criterion-related validity of temperament variables.

The validities in Table 2 are more similar to the results published by Ghiselli in his 1966 book Validity of Occupational Aptitude Tests. Ghiselli, however, summarized validities only for those temperament scales that he evaluated as pertinent to a particular occupational category. We believe that summarizing validities according to constructs enabled us to arrive at conclusions similar to Ghiselli's.

Development of ABLE[1] Scales

The next major task for us was to develop scales that would measure variables identified during the literature review as likely to predict criteria important to the Army. The next page of your handout, List 1, lists the substantive scales we developed for each construct. We developed substantive scales for six constructs: (1) Surgency, (2) Adjustment, (3) Agreeableness, (4) Dependability, (5) Achievement, and (6) Locus of Control.

---

[1] "Assessment of Background and Life Experiences."

We also developed a "Physical Condition" scale to measure physical condition and four response validity scales: (1) Non-Random Response, (2) Social Desirability, (3) Poor Impression, and (4) Self-Knowledge. We developed the "Non-Random Response" scale because we were concerned that some participants would complete the inventory carelessly because the data were gathered for "research purposes only." The "Non-Random Response" scale was developed to detect such inventories. We were also concerned about self-descriptions that were intentionally distorted and wanted to be able to (1) detect intentional distortion, and (2) develop a strategy for dealing with distorted self-descriptions. Thus, we developed the "Social Desirability" scale to detect intentional distortion in an applicant setting (non-draft setting) and "Poor Impression" to detect intentional distortion in a draft setting. We developed a "Self-Knowledge" scale because the literature suggests that people who know themselves well provide more accurate self descriptions, and this greater accuracy moderates the correlation between self description and descriptions or ratings made by others (Gibbons, 1983; Markus, 1983). We hypothesized that "Self-Knowledge" might moderate the relationships between ABLE substantive scales and job performance criteria. In short, we developed four response validity scales to measure accuracy of self-descriptions in order to test the hypothesis that accuracy of self-description moderates the criterion-related validities of ABLE substantive scales.

Evaluation of ABLE Substantive Scales

Once the ABLE temperament scales were developed and pretested, predictor and criterion data were gathered during the summer and fall of 1985 from over 9,000 soldiers. The scale statistics for the temperament inventory, entitled "Assessment of Background and Life Experiences" (ABLE),

appear on the next page of your handout, Table 3. The average number of items in a scale is 15. The median alpha of the substantive scales is .81. Table 4 summarizes the ABLE substantive scale statistics and the correlations of the ABLE substantive scales with each other and with other components of the four-hour predictor battery. As can be seen, the only part of the predictor battery that the ABLE substantive scales correlate with in a substantial way are other ABLE substantive scales. The ABLE substantive scales appear to be tapping a part of the predictor domain not tapped by other measures.

The next page of your handout, Table 5, shows the structure of the ABLE substantive scales. Three factors, Ascendancy, Dependability, and Adjustment, emerged. The scales designed to measure achievement loaded on the same factor as the scale designed to measure Surgency. The literature review indicated that measures of Achievement and Surgency are not highly intercorrelated. Unfortunately, the ABLE scales do not appear to capture the uniqueness of the two constructs.

Criterion-Related Validities. The criterion measures, the development of which was a major part of the research project, were developed by a different part of the research team. The criterion composites, which they also developed, are very briefly described in the next page of your handout, List 2. There are five composites: (1) Core Technical Proficiency, (2) General Soldiering Proficiency, (3) Effort and Leadership, (4) Personal Discipline, and (5) Physical Fitness and Military Bearing. The first two consist mainly of hands-on tests (work samples) and knowledge tests. The other three consist of supervisory and peer ratings and information that can be obtained from personnel records.

The next page of your handout, Table 6, shows the criterion-related

validities of the ABLE scales for these five criteria. The scales are organized according to the literature review taxonomy of temperament constructs. The results suggest that "Achievement" scales are the best predictors of the "Effort and Leadership" criterion; "Dependability" scales are the best predictors of the "Personal Discipline" criterion; and "Physical Condition" is the best predictor of the "Physical Fitness and Military Bearing" criterion, though the "Achievement" scales also correlate with this criterion. These three criteria, which the ABLE substnative scales predict, include the supervisory and peer rating criteria. The other two criteria, which consist of hands-on and knowledge tests, are not predicted with the ABLE substantive scales. The other finding to which I'd like to draw your attention is that, except for "Poor Impression," the response validity scales do not correlate with the supervisory and peer rating criteria. This finding will be relevant later when we analyze the response validity scales in detail.

The next page of your handout, Table 7, shows the criterion-related validities of different types of predictors--cognitive ability, spatial ability, perceptual/psychomotor ability, work environment preferences, temperament, and interests--included in the study. It shows the multiple correlation of each type of predictor with each of the five criteria. As can be seen, the best predictors of the Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing criteria are the ABLE substantive scales. This finding is not surprising, given the literature review and the results that showed that the ABLE substantive scales tap an independent part of the predictor domain.

Evaluation of Response Validity Scales

Recall that we hypothesized that accuracy of self-description moder-

ated the criterion-related validities of the ABLE substantive scales and that we develped four response validity scales to detect four different types of inaccurate self-descriptions.

"Non-Random Response" scale. To evaluate the usefulness of the "Non-Random Response" scale, we examined its moderating effect on the validities of the ABLE substantive scales. We split the sample into two, one group which scored low on "Non-Random Response" scale was designated as "random responders," the remaining sample was designated as "non-random responders." We performed a split-group analysis rather than a moderated regression because the variable of interest had a highly skewed distribution. The results, which are shown in Table 9, indicate that random responding does, indeed, moderate the criterion-related validities of the ABLE substantive scales. Though the validities of the ABLE substantive scales are not uniformly zero for the random responders, typically the validities are significantly lower.

"Self-Knowledge" scale. Previous research, as mentioned earlier, has shown that self-knowledge moderates the correlation between self-description and descriptions or ratings provided by others. We examined the extent to which self-knowledge moderates the relationship between self-descriptions as measured by the ABLE substantive scales and job performance criteria. For each ABLE substantive scale, we computed (1) the zero-order correlation with each criterion, (2) the multiple correlation of the substantive scale and "Self-Knowledge" scale with each criterion, and (3) the multiple correlation based on moderated regression. In moderated-regression analysis, the multiple correlation is incremented by an interaction term, in this case the interaction between "Self-Knowledge" and the particular ABLE substantive scale. We compared these three coefficients for each substantive scale. If Self-Knowledge moderates the criterion-related

validities of the substantive scales, the value of the moderted multiple correlation coefficient would be greater than the multiple correlation. If the values are similar, the relationship is not moderated; a linear model accounts for as much of the variance as the non-linear model. If both values are similar to the zero-order correlation of the substantive scale with the criterion, then the Self-Knowledge scale increments the validity in neither a linear nor a non-linear (moderated) fashion. As is shown in Table 10, the "Self-Knowledge" scale contributes nothing, in either a linear or non-linear way, to the prediction of the criteria.

"Social Desirability" and "Poor Impression" scales. The same logic that applies to the "Non-Random Response" scales applies to the "Social Desirability" and "Poor Impression" scales, that is, accuracy of self-description moderates the relationship between ABLE substantive scales and job criteria. First, though, we wanted to know if "Social Desirability" and "Poor Impression" detected intentional distortion.

To learn whether the "Social Desirability" and "Poor Impression" scales detected intentional distortion, we conducted a faking study with 245 soldiers in which we instructed them to respond honestly, to fake good, and/or to fake bad. Table 11 summarizes the results of that study. Clearly, soldiers were able to distort their self-descriptions when instructed to do so. The median effect size for change in ABLE substantive mean scale scores in the honest and fake good conditions was approximately half a standard deviation. The median effect size for change in ABLE substantive mean scale scores in the honest and fake-bad conditions was over two standard deviations. Fortunately, the "Social Desirability" scale detected faking good--it changed approximately one standard deviation--and the "Poor Impression" scale detected faking bad--it changed over two-and-a-

half standard deviations.

We then examined the extent to which the "Social Desirability" scale moderated the criterion-related validities of the ABLE substantive scales. Unfortunately, we had criterion data for only a few of the soldiers in the faking study. Those data would have been the best to examine because we knew their motivation--they were faking good. We turned to the next best set of data--the concurrent validation sample--to answer the question. Again, our variable had a highly skewed distribution, so we used a split-group technique to investigate the moderating effects of "Social Desirability." We split the group in two. We chose the cutting point to be approximately the mean of the fake good group of the faking study. Thus, the "high" group scored approximately at or above the mean of a group known to be faking. Table 12 shows the results. The "Social Desirability" scale does moderate, slightly, the validities of the ABLE substantive scales. Interestingly, the validities for the "Personal Discipline" criterion are least affected.

Recall that the "Poor Impression" scale correlated with the criteria; thus, it was inappropriate to investigate its moderating effects on the validities of the other ABLE scales, at least in the present sample. We, therefore, examined the contribution of the "Poor Impression" scale as an independent predictor in a linear model. Your next handout, Table 13, shows the zero-order correlation of each of the ABLE substantive scales with each criterion, as well as the multiple correlation of each ABLE substantive scale when "Poor Impression" is included as an independent predictor. As can be seen, the "Poor Impression" scale does increment the validities of the ABLE substantive scales.

Summary

To summarize the results of our work:

1. Constructs provide a strategy to make sense of the literature on the criterion-related validities of temperament scales;
2. The "Big Five" as a taxonomy results in quite a heterogeneous grouping of scales, though they do highlight constructs that are "good bets" for predictors;
3. Temperament variables measure a part of the predictor domain untapped by most other types of predictors;
4. Temperament variables predict certain kinds of criteria, criteria that are not well-predicted by most other predictors;
5. "Self-Knowledge" does not appear to affect the relationship between other temperament variables and job performance criteria; and
6. The response validity scales that detect random responding and self descriptions that are overly positive or negative can be used to increment the validities of temperament variables for job performance criteria.

References

Ghiselli, E. E. (1966). *Validity of occupational aptitude tests.* New York: John Wiley & Sons, Inc.

Gibbons, F. X. (1983). Self-attention and self-report: The "veridicality" hypothesis. *Journal of Personality, 51*, 517-542.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18*, 135-164.

Markus, H. (1983). Self-knowledge: An expanded view. *Journal of Personality, 51*, 543-565.

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.

Tupes, E. C., & Christal, R. E. (1961, May). *Recurrent personality factors based on trait ratings* (ASD-TR-61-97). Lackland Air Force Base, TX: Aeronautical Systems Divisions, Personnel Laboratory.

Table 1 Mean Within-Category and Between Category Correlations of Temperament Scales

	Surgency	Adjustment	Agreeableness	Dependability	Intellectance	Affiliation	Miscellaneous
Surgency	Mean$_r$=(.46) SD$_r$=.16 N$_r$=146						
Adjustment	Mean$_r$=.20 SD$_r$=.18 N$_r$=321	Mean$_r$=(.43) SD$_r$=.19 N$_r$=165					
Agreeableness	Mean$_r$=.04 SD$_r$=.17 N$_r$=173	Mean$_r$=.24 SD$_r$=.16 N$_r$=162	Mean$_r$=(.37) SD$_r$=.14 N$_r$= 44				
Dependability	Mean$_r$=-.08 SD$_r$= .16 N$_r$= 286	Mean$_r$=.13 SD$_r$=.20 N$_r$=276	Mean$_r$=.06 SD$_r$=.17 N$_r$=166	Mean$_r$=(.34) SD$_r$=.18 N$_r$=121			
Intellectance	Mean$_r$=.12 SD$_r$=.15 N$_r$=175	Mean$_r$=.02 SD$_r$=.14 N$_r$=193	Mean$_r$=.04 SD$_r$=.16 N$_r$= 94	Mean$_r$=-.12 SD$_r$= .18 N$_r$= 162	Mean$_r$=(.40) SD$_r$=.19 N$_r$= 52		
Affiliation	Mean$_r$=.09 SD$_r$=.21 N$_r$=157	Mean$_r$=.00 SD$_r$=.16 N$_r$=150	Mean$_r$=.10 SD$_r$=.17 N$_r$= 98	Mean$_r$=.08 SD$_r$=.14 N$_r$=160	Mean$_r$=-.14 SD$_r$= .15 N$_r$= 84	Mean$_r$=(.33) SD$_r$=.16 N$_r$= 45	
Miscellaneous	Mean$_r$=.09 SD$_r$=.17 N$_r$=392	Mean$_r$=.12 SD$_r$=.18 N$_r$=419	Mean$_r$=.02 SD$_r$=.18 N$_r$=215	Mean$_r$=.02 SD$_r$=.18 N$_r$=361	Mean$_r$=.04 SD$_r$=.17 N$_r$=242	Mean$_r$=-.04 SD$_r$= .15 N$_r$= 208	Mean$_r$=(.05) SD$_r$=.20 N$_r$=246

Table 2 Meta Analysis of Criterion-Related Validity Studies [1] That Used Temperament Predictors

	Criterion											
	Educational		Training		Job Involvement		Job Proficiency		Negative Adjustment: Delinquency		Negative Adjustment: Substance Abuse	
Predictor Construct[2]	Number Predictors	mean r	Number Predictors	mean r	Number Predictors	mean r	Number Predictors	mean r	Number Predictors	mean r	Number Predictors	mean r
*Surgency	42	.15	47	.08	21	.04	175	.04	8	[-.29]	30	.06
Affiliation	5	-.04	0	---	4	.06	16	-.01	0	---	4	-.03
*Adjustment	44	[.26]	44	[.16]	21	[.13]	146	[.13]	10	[-.43]	31	-.07
*Agreeableness	9	.01	5	.10	4	.02	48	-.01	1	[-.31]	8	-.04
*Dependability	24	[.15]	26	.11	18	[.17]	102	[.13]	10	[-.27]	25	[-.28]
*Intellectance	6	[.18]	7	.14	8	-.10	32	.01	1	[-.24]	2	[.18]
Achievement	8	[.30]	4	[.33]	4	[.24]	0	---	4	[-.35]	0	---
Masculinity	8	[-.16]	3	.09	10	.10	0	---	3	.02	8	[-.18]
Locus of Control	1	.32	2	[.29]	7	[.25]	0	---	0	---	0	---

[1] Time Period 1960-1984.

[2] A star denotes the construct is one of the "Big Five" constructs.

Note: Correlations are not corrected for unreliability or range restrictions.

List 1

ABLE[1] Scales Organized According to Construct Intended to Measure

SUBSTANTIVE SCALES:

Surgency
- Dominance
- Energy Level

Adjustment
- Emotional Stability

Agreeableness (Likeability)
- Cooperativeness

Dependability
- Nondelinquency
- Traditional Values
- Conscientiousness

Achievement
- Work Orientation
- Self Esteem

Locus of Control
- Internal Control

Physical Condition
- Physical Condition

RESPONSE VALIDITY SCALES:

- Non-Random Response
- Social Desirability
- Poor Impression
- Self-Knowledge

[1] Inventory developed by PDRI for the Army Research Institute entitled "Assessment of Background and Life Experience."

Table 3 ABLE Scale Statistics for Total Group[1]
(Concurrent Sample; Revised Trial Battery)

	No. Items	N	Mean	S.D.	Internal Consistency Reliability (Alpha)	Test-Retest Reliability[2]
ABLE SUBSTANTIVE SCALES						
Emotional Stability	17	8522	39.0	5.45	.81	.74
Self-Esteem	12	8472	28.4	3.70	.74	.78
Cooperativeness	18	8494	41.9	5.28	.81	.76
Conscientiousness	15	8504	35.1	4.31	.72	.74
Nondelinquency	20	8482	44.2	5.91	.81	.80
Traditional Values	11	8461	26.6	3.72	.69	.74
Work Orientation	19	8498	42.9	6.06	.84	.78
Internal Control	16	8485	38.0	5.11	.78	.69
Energy Level	21	8488	48.4	5.97	.82	.78
Dominance	12	8477	27.0	4.28	.80	.79
Physical Condition	6	8500	14.0	3.04	.84	.85
ABLE RESPONSE VALIDITY SCALES						
Social Desirability	11	8511	15.5	3.04	.63	.63
Self-Knowledge	11	8508	25.4	3.33	.65	.64
Non-Random Response[3]	8	9188	7.4	1.19	—	.30
Poor Impression	23	8492	1.5	1.85	.63	.61

[1] Total group after screening for missing data and random responding.
[2] N = 408 - 412 for test-retest correlations (N = 414 for Non-Random Response test-retest correlations).
[3] Screened only for missing data.

Table 4 ABLE Substantive Scales: Summary

(Revised Trial Battery)

	Range	Median
Reliability:		
Internal Consistency (Alpha)	.69 - .84	.81
Test-Retest	.69 - .85	.78
Relationship to Predictor Variables:		
Correlation ABLE Substantive Scales	.00 - .73	.30
Correlation Interest Scales	.00 - .43	.09
Correlation Preferred Work Environment Scales	.00 - .35	.13
Correlation Perceptual/Psychomotor Measures	.00 - .13	.03
Correlation Cognitive Measures	.00 - .20	.05
ASVAB[1] Adj. R^2	.01 - .04	.01

[1] Mental ability test currently used by military.

Table 5

ABLE Substantive Scales: Factor Analysis[1]

(Concurrent Sample; Revised Trial Battery)

	Factor I Ascendancy	Factor II Dependability	Factor III Adjustment	h^2
Dominance	[.84]	.04	.17	.73
Self-Esteem	[.77]	.13	.37	.75
Work Orientation	[.72]	.47	.15	.77
Energy Level	[.66]	.32	.47	.76
Traditional Values	.13	[.84]	.10	.73
Nondelinquency	.04	[.82]	.26	.74
Conscientiousness	.49	[.72]	.07	.76
Internal Control	.27	.47	.44	.48
Emotional Stability	.33	.04	[.85]	.84
Cooperativeness	.17	.46	[.67]	.69
				7.25

[1] Principal component analysis, varimax rotation.

Note: N = 8367

List 2

Criterion Composites[1]

Core Technical Proficiency - a) hands-on tests of MOS-specific technical knowledge and skills; and b) tests of school and job knowledge.

General Soldiering Proficiency - a) hands-on tests of general soldiering skill; and b) general soldiering knowledge and skill test items.

Effort & Leadership - a) supervisory and peer ratings of effort and leadership, overall effectiveness, MOS effectiveness and predicted combat effectiveness; and b) letters and certificates of commendation and other achievements.

Personal Discipline - a) supervisory and peer ratings of personal control and discipline; and b) disciplinary actions and other negative indicators in personnel files.

Physical Fitness & Military Bearing - a) supervisory and peer ratings of physical fitness and military bearing; and b) physical readiness tests.

[1]Data gathered at same time as Trial Battery was administered, i.e., summer and fall of 1985.

Table 6 Validities of ABLE Scales for Job Performance Criteria: Zero-Order Correlations (Revised Trial Battery; Concurrent Validity Study)

Predictor	Criterion: Core Technical Proficiency	General Soldiering Proficiency	Effort & Leadership	Personal Discipline	Physical Fitness & Military Bearing
Surgency:					
. Dominance	.01	.01	.15	.02	.18
Achievement:					
. Self Esteem	.02	.01	[.20]	.13	[.20]
. Work Orientation	.02	.02	[.23]	.18	[.21]
. Energy Level	.02	.02	[.22]	.14	[.25]
Adjustment:					
. Emotional Stability	.02	.02	.17	.12	.16
Agreeableness (Likeability)					
. Cooperativeness	.01	.02	.15	[.21]	.14
Dependability:					
. Traditional Values	.03	.06	.13	[.25]	.16
. Non-delinquency	.05	.07	.12	[.29]	.14
. Conscientiousness	.02	.02	.18	[.23]	.22
Others:					
. Internal Control	.04	.05	.13	.13	.13
. Physical Condition	-.04.	-.05	.09	-.03	[.29]
Response Validity Scales:					
. Non-Random Response[1]	.13	.14	.07	.10	.02
. Social Desirability	-.07	-.06	.02	.05	.07
. Poor Impression	-.04	-.05	-.15	-.15	-.16
. Self-Knowledge	-.04	-.03	.07	.05	.13

[1]Correlations are based on unscreened data for this scale. N varies from 8424 to 9322 for this scale.

Note: N varies from 7666 to 8477.

Note: A box indicates notable predictor/criterion construct relationships.

Table 7

Multiple Correlations[1] of Six Independent Predictor Composites with each of Five Job Performance Criteria

(Concurrent Validity Study)

Predictor Composites	Criterion Composites: Core Technical Proficiency	General Soldiering Proficiency	Effort & Leadership	Personal Discipline	Physical Fitness & Military Bearing
ASVAB[2] (mental ability test)	[.62]	[.64]	[.35]	.20	.14
Spatial Abilities	.56	.62	.26	.14	.11
Perceptual/Psychomotor Abilities (computerized)	.54	.58	.30	.12	.10
Work Environment Preferences	.28	.27	.20	.10	.11
Temperament (and physical activities scale)	.26	.24	[.34]	[.33]	[.36]
Interests	.34	.34	.26	.14	.13

[1]Multiple Rs are adjusted for shrinkage and corrected for restriction in range, but not corrected for criterion unreliability.

[2]Mental ability test currently used by military.

Note: Entries in table are averaged across 9 Army military occupational specialties (MOS) with complete criterion data. Total sample is 3902. Sample sizes range from 281 to 570; median = 432.

Note: Boxes denote the two best predictors of the criterion space.

Table 9 Moderating Effects of Random Responding on Correlations Between ABLE Scales and Job Performance Criteria

ABLE SCALE	CRITERION: Effort/Leadership Low (Random)	Effort/Leadership High (Non-Random)	Personal Discipline Low (Random)	Personal Discipline High (Non-Random)	Physical Fitness/Bearing Low (Random)	Physical Fitness/Bearing High (Non-Random)
Surgency:						
Dominance	.06	.15	.05	.02	.18	.18
Achievement:						
Self-Esteem	-.00	.15	.03	.09	.08	.18
Work Orientation	.05	.23	.08	.18	.12	.21
Energy Level	.07	.22	.10	.14	.20	.25
Adjustment:						
Emotional Stability	.11	.17	.08	.12	.09	.16
Agreeableness:						
Cooperativeness	.13	.15	.17	.21	.10	.14
Dependability:						
Traditional Values	.07	.13	.19	.25	.18	.16
Nondelinquency	.09	.12	.22	.29	.14	.14
Conscientiousness	.05	.18	.11	.23	.16	.22
Others:						
Internal Control	.00	.13	.03	.13	.05	.13
Physical Condition	-.03	.09	-.00	-.03	.16	.29

N ranges from 659 to 675 for group scoring low on "Non-Random Response" scale
N ranges from 8336 to 8477 for group scoring high on "Non-Random Response" scale
Note: Statistically significant differences at $P \leq .05$ is approximately .04.

Table 10 Incremental Validities of ABLE Scales When "Self-Knowledge" Scale Is Included in Predictor Equation (Linear and Non-Linear Models)

ABLE SCALE	CRITERION: Effort/Leadership r	R	R_{mod}	Personal Discipline r	R	R_{mod}	Physical Fitness/Bearing r	R	R_{mod}
Surgency:									
Dominance	.15	.15	.15	.02	.04	.04	.18	.19	.20
Achievement:									
Self-Esteem	.20	.20	.20	.12	.12	.12	.20	.21	.22
Work Orientation	.23	.23	.23	.18	.18	.18	.21	.22	.22
Energy Level	.22	.22	.22	.14	.14	.14	.25	.26	.26
Adjustment:									
Emotional Stability	.17	.18	.18	.12	.12	.13	.16	.20	.20
Agreeableness:									
Cooperativeness	.15	.15	.15	.21	.21	.21	.14	.17	.17
Dependability:									
Traditional Values	.14	.14	.14	.25	.25	.25	.17	.19	.19
Nondelinquency	.13	.13	.13	.29	.29	.29	.14	.17	.17
Conscientiousness	.18	.18	.18	.23	.23	.23	.22	.22	.22
Others:									
Internal Control	.13	.13	.13	.13	.13	.13	.13	.16	.16
Physical Condition	.09	.09	.09	.03	.05	.05	.29	.30	.30

N ~ 8440

Note: The r column is the zero-order correlation of the substantive scale with the criterion.

The R column is the multiple correlation of the substantive scale and Self-Knowledge with the criterion, i.e., Self-Knowledge is included as an independent predictor.

The R_{mod} is the multiple correlation based on moderated regression. It increments the multiple correlation by including an interaction term. A difference betwen R and R_{mod} suggests that Self-Knowledge moderates the validity.

Table 11 Effect Size of Differences Between Honest and Faking Conditions for ABLE Response Validity Scales and Substantive Scales

	EFFECT SIZE	
	Honest vs. Fake Good	Honest vs. Fake Bad
PREDICTOR SCALE		
ABLE Substantive Scales	-.49	2.10
ABLE Response Validity Scales:		
Social Desirability	-1.02	-.53
Non-Random Response	.45	3.16
Poor Impression	-.09	-2.67

Table 12 Moderating Effects of "Social Desirability" Scale on Correlations Between ABLE Scales and Job Performance Criteria

ABLE SCALE	CRITERION: Effort/Leadership Non-High[1]	Effort/Leadership High[2]	Personal Discipline Non-High[1]	Personal Discipline High[2]	Physical Fitness/Bearing Non-High[1]	Physical Fitness/Bearing High[2]
Surgency:						
Dominance	.15	.14	.00	.06	.18	.17
Achievement:						
Self-Esteem	.21	.18	.12	.12	.21	.17
Work Orientation	.25	.20	.17	.16	.22	.17
Energy Level	.23	.20	.13	.15	.27	.20
Adjustment:						
Emotional Stability	.17	.16	.11	.12	.16	.13
Agreeableness:						
Cooperativeness	.16	.13	.20	.21	.14	.12
Dependability:						
Traditional Values	.14	.11	.26	.22	.18	.11
Nondelinquency	.13	.12	.28	.29	.14	.11
Conscientiousness	.19	.14	.22	.22	.24	.14
Others:						
Internal Control	.13	.12	.12	.15	.15	.08
Physical Condition	.08	.09	-.03	-.02	.28	.29

[1] N ranges from 2428 to 2480 for group scoring high on "Social Desirability" scale

[2] N ranges from 5896 to 5997 for group scoring Non-High on "Social Desirability" scale

Note: A statistically significant difference at $p \leq 05$ is approximately .03

Table 13 Incremental Validities of ABLE Scales When "Poor Impression" Scale is Included in Predictor Equation (Linear Model)

ABLE SCALE	CRITERION: Effort/Leadership r	Effort/Leadership R	Personal Discipline r	Personal Discipline R	Physical Fitness/Bearing r	Physical Fitness/Bearing R
Surgency:						
Dominance	.15	.19	.02	.15	.18	.22
Achievement:						
Self-Esteem	.20	.22	.12	.17	.20	.22
Work Orientation	.23	.25	.18	.20	.21	.23
Energy Level	.22	.22	.14	.17	.25	.26
Adjustment:						
Emotional Stability	.17	.18	.12	.16	.16	.18
Agreeableness:						
Cooperativeness	.15	.18	.21	.22	.14	.17
Dependability:						
Traditional Values	.14	.17	.25	.26	.17	.20
Nondelinquency	.13	.17	.29	.29	.14	.18
Conscientiousness	.18	.20	.23	.24	.22	.23
Others:						
Internal Control	.13	.17	.13	.17	.13	.17
Physical Condition	.09	.16	.03	.16	.29	.31

N ~ 8400

Note: A statistically significant difference at $p \leq .05$ is approximately .02

PROJECT A VALIDITY RESULTS:

THE RELATIONSHIP BETWEEN PREDICTOR AND CRITERION DOMAINS

Jeffrey J. McHenry
American Institutes for Research

Leaetta M. Hough
Jody L. Toquam
Mary Ann Hanson
Steven Ashworth
Personnel Decisions Research Institute

Presented at the Annual Conference of the
Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Project A Validity Results: The Relationship between Predictor and Criterion Domains

The purpose of this paper is to describe the relationship between the predictor scores described in the Peterson, Hough, Dunnette, Rosse, Houston, Toquam, and Wing (1987) and the criterion scores described in the Campbell, Harris, McHenry, and Arabian (1987) paper.

This paper includes five parts. In the first part, we describe the creation of predictor composite scores from the predictor test and scale scores described in Peterson et al. (1987). In the second part, we show the relationship between the predictor composite scores within each predictor domain and the five job performance constructs described by Campbell et al. (1987). In the third part, we demonstrate how the new predictor tests increment the validity of the Army's current selection battery, the Armed Services Vocational Aptitude Battery (ASVAB). In the fourth part, we describe the relationship between the new predictor tests and two "method factors" that we identified in our analyses of the job performance measures. Finally, in the fifth part, we discuss how the predictor-criterion relationships uncovered in the validity analyses contribute to the understanding of job performance in the Army.

Formation of Predictor Composites

The preliminary analyses of the new Project A predictor tests indicated that 65 reliable predictor scores could be computed from the six spatial tests, the 10 computer tests, and the temperament/personality, vocational interest, and job reward preference inventories (Peterson et al., 1987). In addition, scores from the nine ASVAB subtests were available from Army records. Table 1 shows how these predictor scores were distributed among various domains within the predictor space. The ASVAB subtests measured nine cognitive abilities. The spatial tests measured six different aspects of spatial ability. The ten computer tests yielded 20 measures of perceptual-psychomotor abilities. The ABLE provided measures of 11 temperaments/personality traits. The AVOICE assessed 22 vocational interests. Finally, the JOB measured six types of job reward preferences.

There were several problems that precluded using these 74 scores directly in the Project A validity analyses. First, as Table 2 shows, the number of subjects with complete predictor and criterion data within the nine target Project A jobs ranged from 289 for Single Channel Radio Operator to 597 for Military Police (Young, Harris, Hoffman & Houston, 1987). Even for Military Police, the ratio of subjects to variables was only 8:1. Our intent was to use multiple regression to estimate the correlation between the predictors and job performance constructs. This ratio is far less than the ratio of 10:1 that many statisticians say is the minimum required to obtain stable estimates of multiple regression coefficients and the coefficient of multiple correlation $\underline{R}$. Since we were faced with a fixed number of subjects, the only way to improve this ratio was to reduce the number of predictor scores.

Second, scores from many of the predictor tests were highly

Table 1

Assessment of the Predictor Space

Predictor Domain	Measure[a]	Number of Test or Scale Scores	Number of Composite Scores
General Cognitive Ability	Armed Services Vocational Aptitude Battery (ASVAB)	9 Subtest Scores	4 Composite Scores
Spatial Ability	Spatial Test Battery	6 Test Scores	1 Composite Score
Perceptual-Psychomotor Abilities	Computer Battery	20 Test Scores	6 Composite Scores
Temperament/ Personality	Assessment of Background and Life Experiences (ABLE)	11 Scale Scores[b]	4 Composite Scores
Vocational Interests	Army Vocational Interest Career Examination (AVOICE)	22 Scale Scores	6 Composite Scores
Job Reward Preferences	Job Orientation Blank (JOB)	6 Scale Scores	3 Composite Scores

[a]All measures except the ASVAB were developed specifically for Project A.
[b]The ABLE included 4 additional response validity scales.

Table 2

The Number of Incumbents in the Nine Army Enlisted Jobs Studied

Enlisted Job	Number of Incumbents
Infantryman	491
Cannon Crewmember	464
Armor Crewman	394
Single Channel Radio Operator	289
Light Wheel Vehicle Mechanic	478
Motor Transport Operator	507
Administrative Specialist	427
Medical Specialist	392
Military Police	597

intercorrelated. For example, the average intercorrelation among the six Project A spatial tests was .46. This multicollinearity results in unstable estimates of multiple regression coefficients. This situation can be remedied by combining the correlated test scores into a single composite. To the extent that the tests are highly intercorrelated, the composite score should contain all of the reliable variance included in any of the individual test scores. Also, the composite should be more reliable than any of the individual test scores, since it will be based on more items than any the score from any single test.

Because of these two problems, the 74 predictor test and scale scores were combined into 20 predictor composites before predictor-criterion relationships were explored. With one exception (which will be noted below), these composites were formed simply by summing standardized test or scale scores; that is, in all instances but one, unit weights were used to compute composite scores from test and scale scores.

Three principles were used to guide the formation of composite scores. First, we attempted to keep the number of composites to a minimum. We expected that this would increase the stability of all of the multivariate statistics we intended to compute in exploring predictor-criterion relationships. Second, we sought to maintain homogeneity or internal consistency within composites. To guide in this effort, we studied the intercorrelations among test or scale scores. We also used principal components analysis to identify tests or scales with similar patterns of factor loadings. Test or scale scores with reasonably high intercorrelations and similar patterns of factor loadings tended to be grouped into the same composite. We believed that this would eliminate any problems associated with predictor multicollinearity. Third, even if we found that two or more test or scale scores were reasonably highly correlated and had similar patterns of factor loadings, we grouped them into the same composite only if we expected that they would have similar patterns of correlations with our job performance constructs. Expert judgments of expected predictor-criterion relationships were available to direct us in this task (Wing, Peterson & Hoffman, 1984).

Figure 1 shows how the nine ASVAB subtests were combined into four composite scores. The four composites were Technical, Quantitative, Verbal, and Speed. In computing the Technical composite score, the Electronics Information subtest received a weight of one-half, while the Mechanical Comprehension and Auto Shop subtests received unit weights. The weight for the Electronics Information subtest was only one-half because a factor analysis indicated that the loading of the Electronics Information on the Technical factor of the ASVAB was only about one-half as large as the loading of the Mechanical Comprehension and Auto Shop subtests.

As noted above, the six spatial tests were all highly intercorrelated. Therefore, as Figure 2 shows, these six tests were combined into a single composite score.

Six composite scores were computed from the 20 perceptual-psychomotor test scores from the computer battery. These six composites were Psychomotor, Complex Perceptual Speed, Complex Perceptual Accuracy, Number Speed and Accuracy, Simple Reaction Speed, and Simple Reaction Accuracy. Figure 3 shows how the 20 test scores were combined into these six composites.

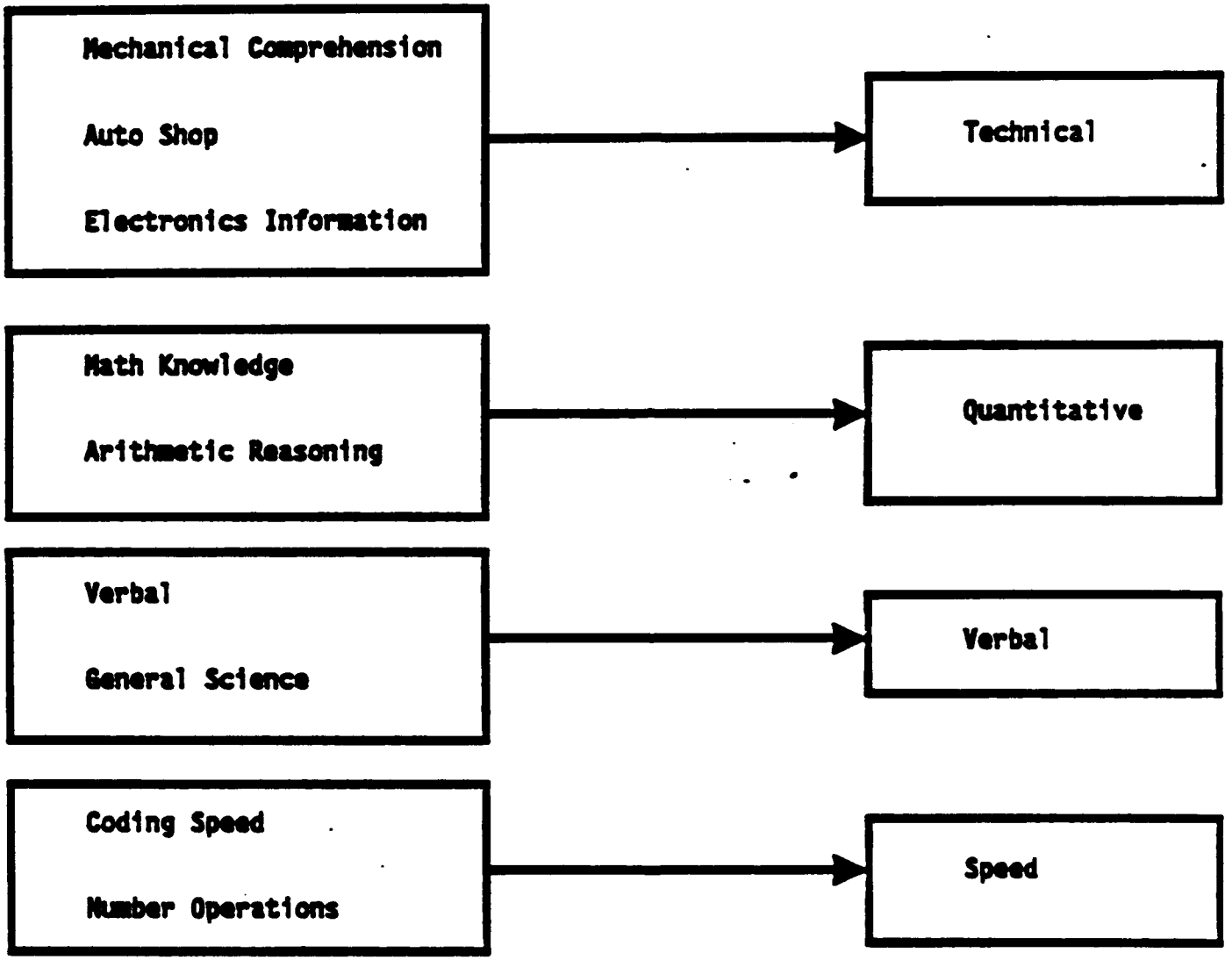


Figure 1. Formation of general cognitive ability composites from ASVAB subtests.

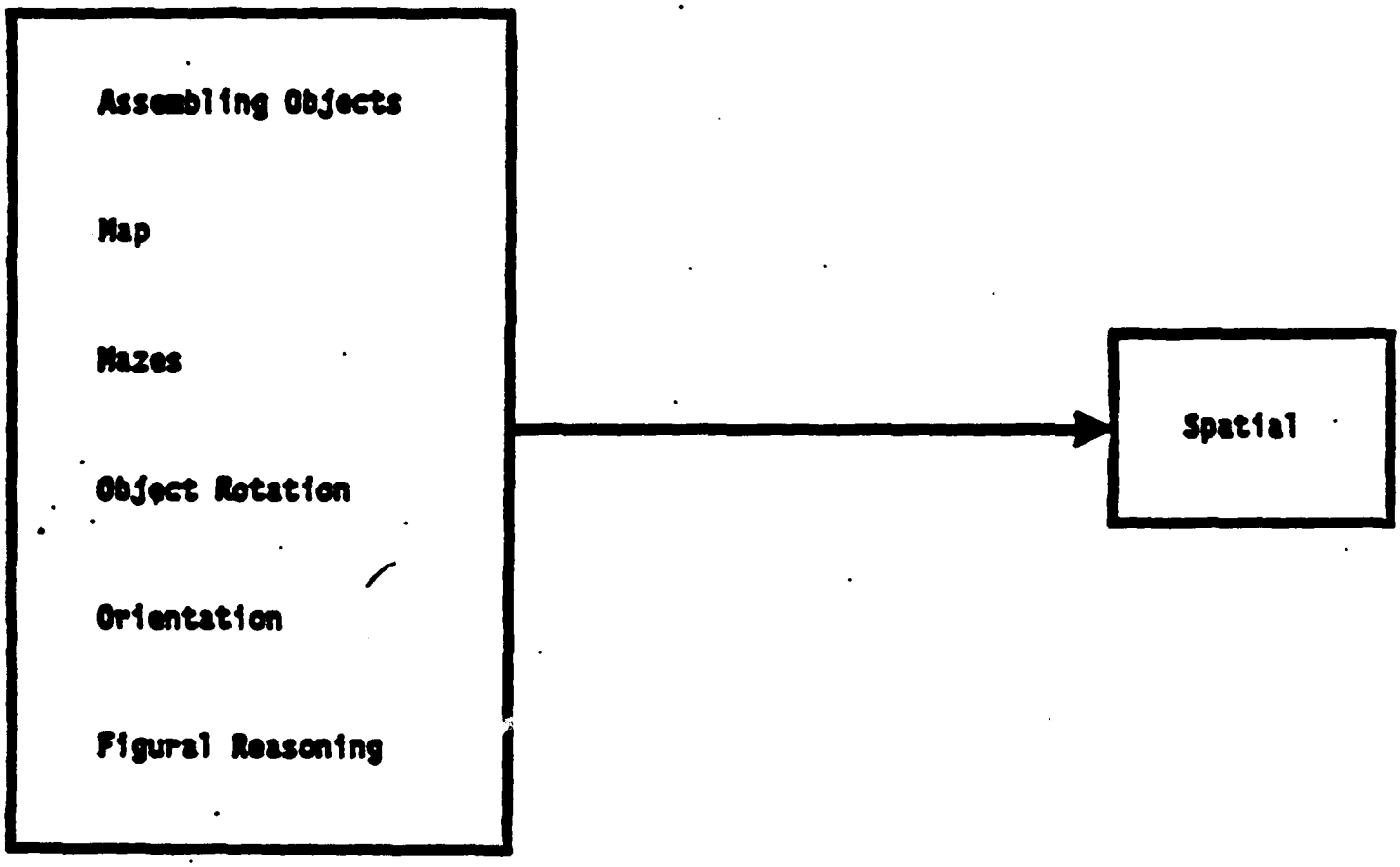


Figure 2. Formation of spatial ability composite from spatial battery test scores.

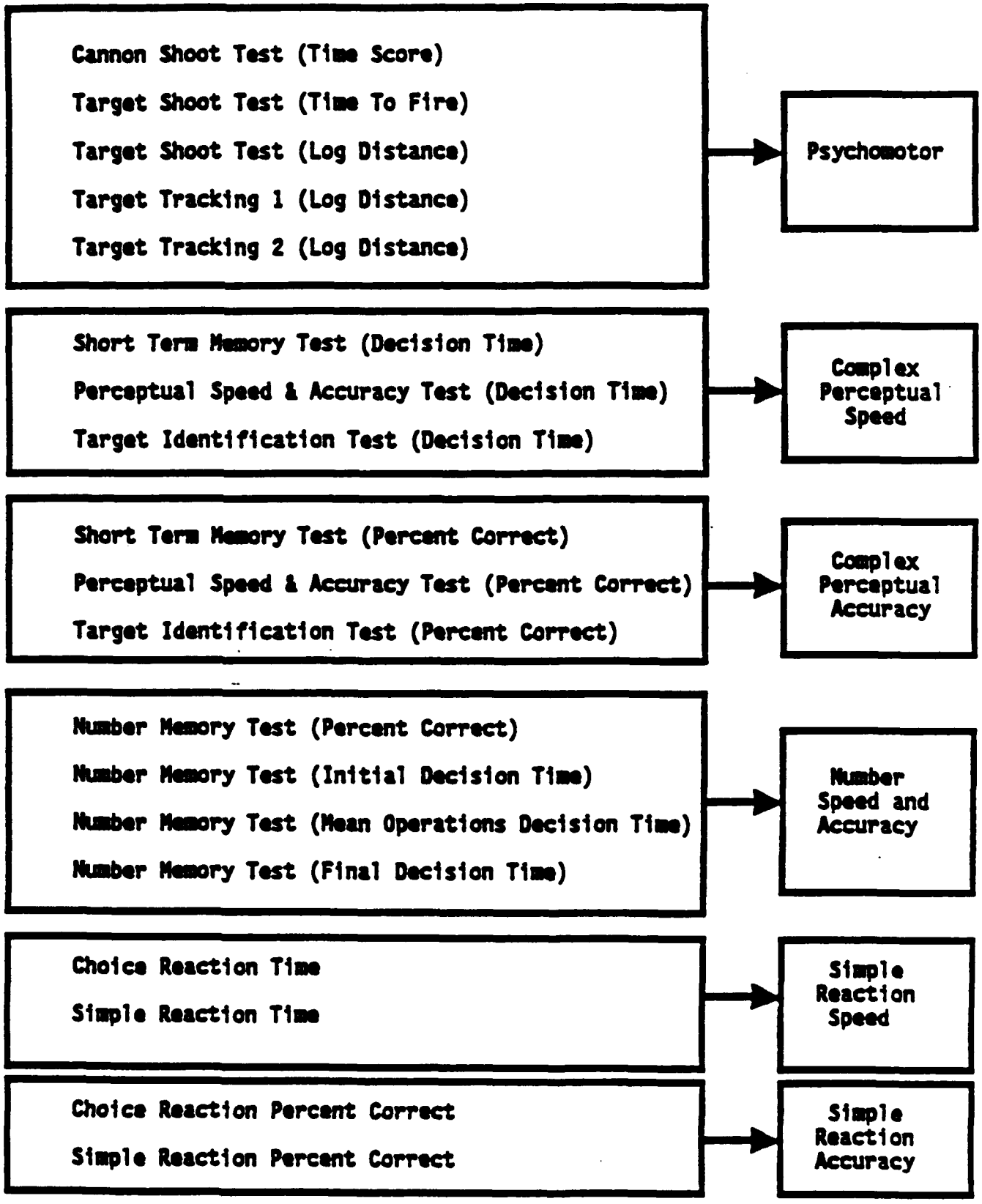


Figure 3. Formation of perceptual-psychomotor ability composites from computer battery test scores.

Four temperament/personality composites were computed from the ABLE scales (see Figure 4). The composites included Achievement Orientation, Dependability, Adjustment, and Physical Condition. Four of the 11 ABLE scales were not included in any composite.

Figure 5 shows that six vocational interest composites were computed from the 22 AVOICE scales. These composites were Skilled Technical, Structural/Machines, Combat-Related, Audiovisual Arts, Food Service, and Protective Services.

Finally, the six scales of the JOB were combined into three composites (see Figure 6). These were Organizational and Co-Worker Support, Routine Work, and Job Autonomy.

Relationships between Predictor Domains and Job Performance Constructs

Job Performance Constructs

The performance criteria used for this study were the five job performance constructs described in the Campbell et al. (1987) paper. Table 3 provides complete definitions of these five constructs. The first construct, Core Technical Proficiency, refers to a soldier's performance on those tasks that are central to the soldier's job. The second construct, General Soldiering Proficiency, represents a soldier's performance on tasks that are required of all soldiers, regardless of their assigned job. These first two constructs represent the "can do" portion of the job performance space. The third performance construct is Effort and Leadership. This construct reflects the degree to which a soldier tries hard on the job, even under adverse or hazardous conditions, and provides support and encouragement for peers. The fourth construct, Personal Discipline, represents the degree to which a soldier follows Army rules and regulations, maintains high standards of personal conduct, and avoids disciplinary problems. The fifth construct, Physical Fitness and Military Bearing, represents the degree to which a soldier maintains an appropriate military appearance and bearing and stays in good physical condition. These final three performance constructs -- Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing -- represent the "will do" portion of the job performance space, though Effort and Leadership also includes some elements of "can do" performance.

Hypothesized Relationships between Predictor Domains and Job Performance Constructs

Figure 7 depicts the expected relationships between the predictor domains and the five job performance constructs. From the cognitive portion of the predictor space, four ASVAB composite scores were available for general cognitive ability, a spatial battery composite score was available for spatial ability, and six computer battery composite scores were available for perceptual-psychomotor ability. It was hypothesized that these cognitive predictor composite scores would be useful for predicting scores on the two "can do" performance constructs, Core Technical

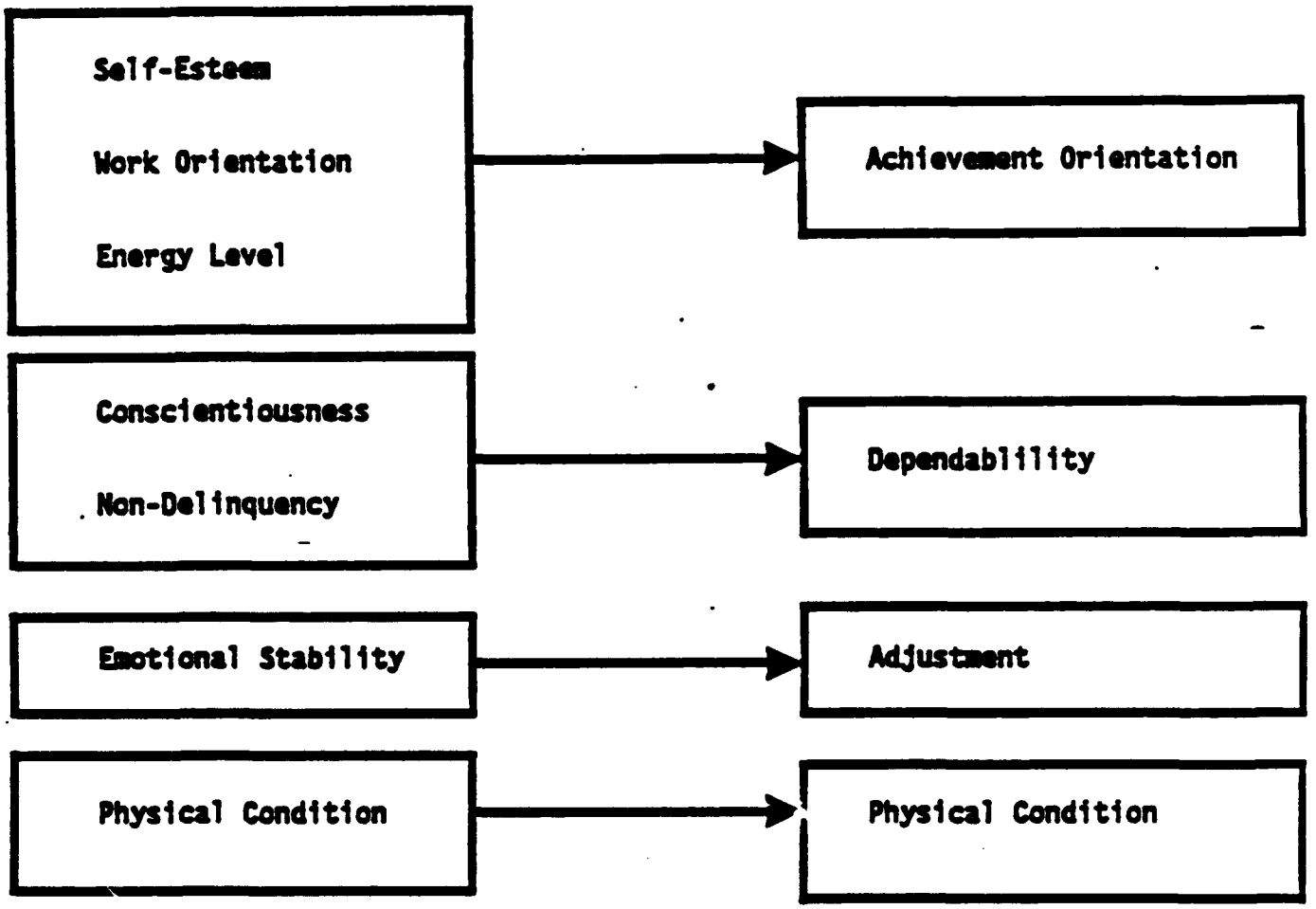


Figure 4. Formation of temperament/personality composites from ABLE scale scores. Four ABLE scales were not used in computing composite scores. These were Dominance, Traditional Values, Cooperativeness, and Internal Control.

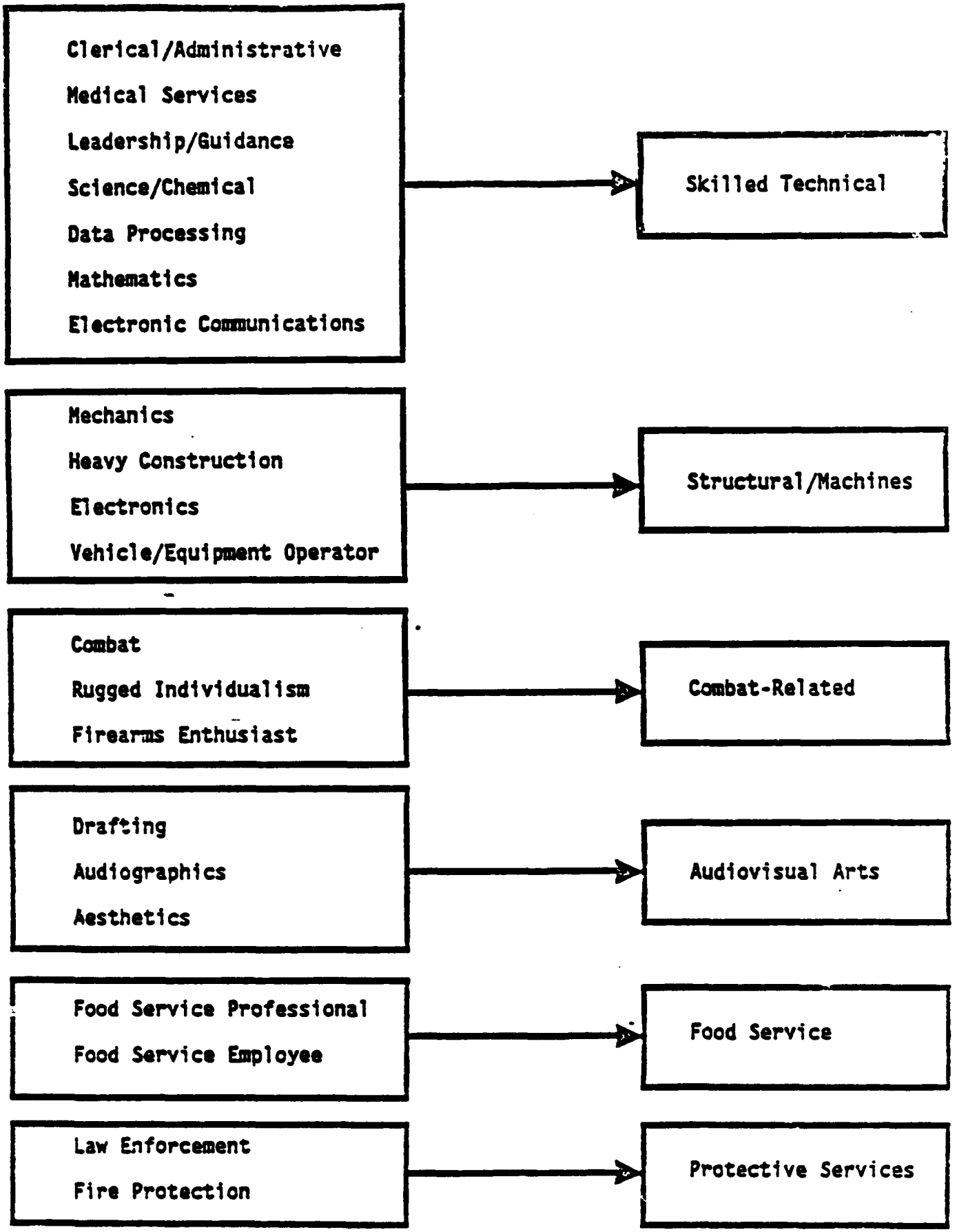


Figure 5. Formation of vocational interest composites from AVOICE scale scores.

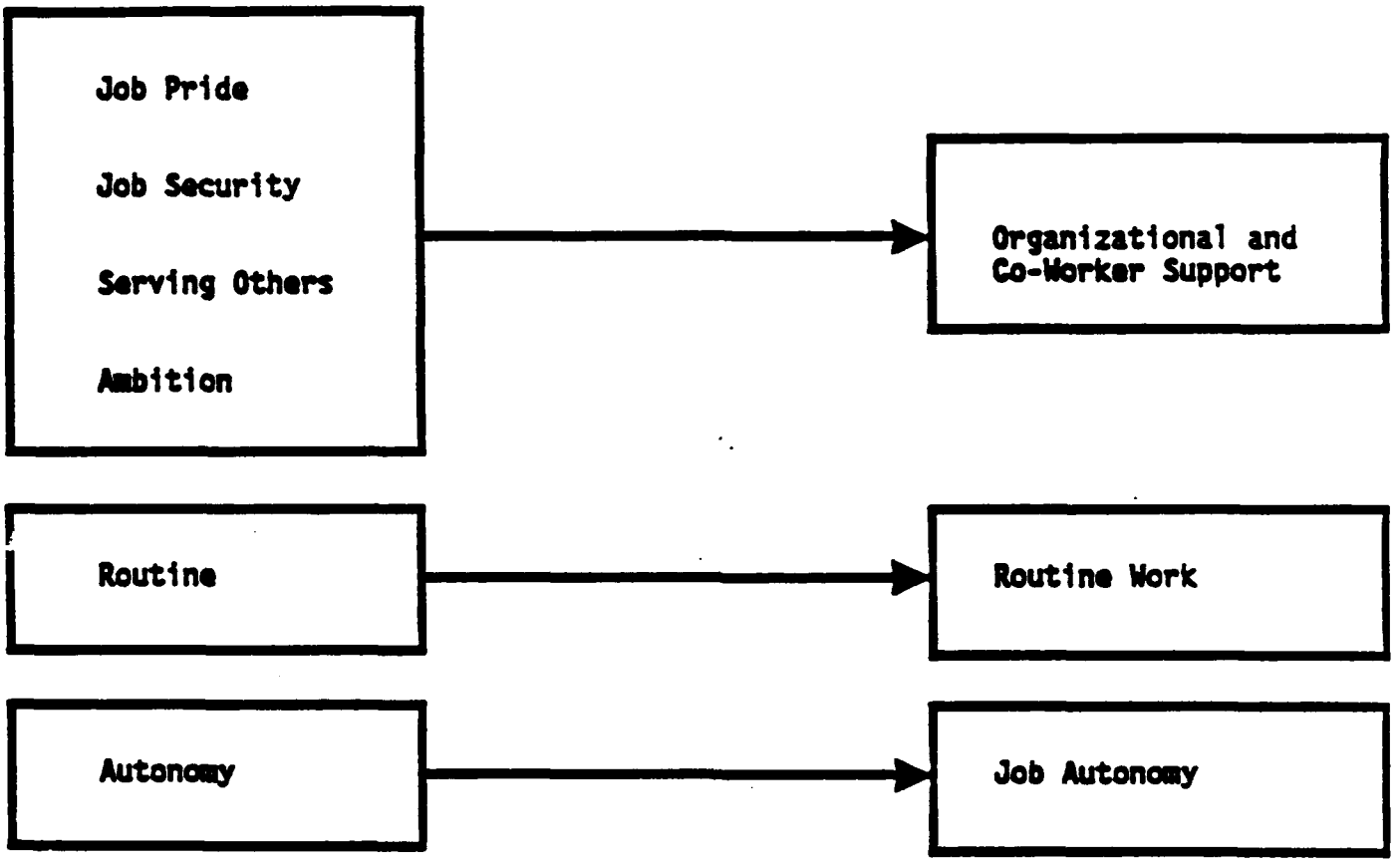


Figure 6. Formation of job reward preference composites from JOB scale scores.

Table 3

Definitions of the Job Performance Constructs

Core Technical Proficiency
This performance construct represents the proficiency with which the soldier performs the tasks that are "central" to the job. The tasks represent the core of the job and they are the primary definers of the job. For example, the first tour Armor Crewman starts and stops the tank engines; prepares the loader's station; loads and unloads the main gun; boresights the M60A3; engages targets with the main gun; and performs misfire procedures. This performance construct does not include the individual's willingness to perform the task or the degree to which the individual can coordinate efforts with others. It refers to how well the individual can execute the core technical tasks the job requires, given a willingness to do so.

General Soldiering Proficiency
In addition to the core technical content specific to a job, individuals in every job also are responsible for being able to perform a variety of general soldiering tasks (e.g., determines grid coordinates on military maps; puts on, wears and removes M17 series protective mask with hood; determines a magnetic azimuth using a compass; collects/reports information -- SALUTE; and recognizes and identifies friendly and threat aircraft). Performance on this construct represents overall proficiency on these general soldiering tasks. Again, it refers to how well the individual can execute general soldering tasks, given a willingness to do so.

Effort and Leadership
This performance construct 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. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient? While appropriate knowledges and skills are necessary for successful performance, this construct is only meant to reflect the individual's willingness to do the job required and to be cooperative and suportive with other soldiers.

Personal Discipline
This performance construct reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems. People who rank high on this construct show a commitment to high standards of personal conduct.

Physical Fitness and Military Bearing
This performance construct represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

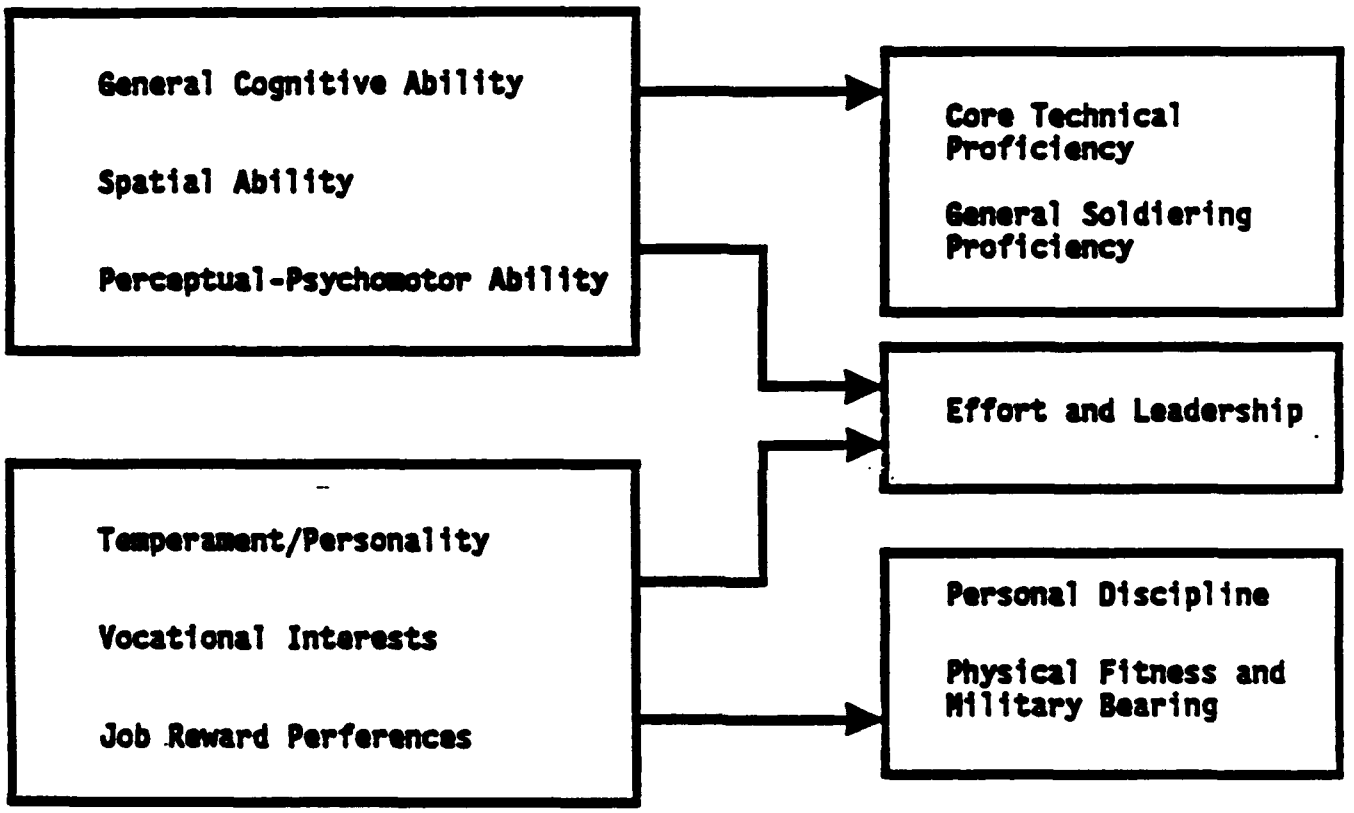


Figure 7. Hypothesized predictor-criterion relationships.

Proficiency and General Soldiering Proficiency. It was hypothesized that the cognitive predictor composite scores also would be useful for predicting scores on Effort and Leadership, since Effort and Leadership also contained some components of "can do" performance.

The four ABLE temperament/personality composite scores, the six AVOICE vocational interest composite scores, and the three job reward preference composite scores from the JOB all were intended to serve as measures of the non-cognitive portion of the predictor space. It was hypothesized that these predictor composites would be most useful for predicting the "will do" job performance constructs, including Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing.

Assessing the Relationships between Predictor Domains and Job Performance Constructs

Statistical procedures. To assess the relationships between predictor domains and job performance constructs, multiple linear regression was used to determine the multiple correlation R of the predictor composites within each domain with each of the five job performance constructs. This was one separately for each of the nine jobs. Each R was corrected for range restriction and adjusted for shrinkage.

The procedure used to correct R for range restriction is one described in Lord and Novick (1968). The procedure adjusts the intercorrelations among the ASVAB subtests so that they match the intercorrelations obtained in a 1980 youth population (Mitchell & Hanser, 1984). The correlations among the predictor composite scores and the performance construct scores are then adjusted according to their correlations with the ASVAB subtests. This means that the correction procedure takes into account any range restriction related to the abilities measured in the ASVAB. However, it fails to consider factors that may reduce the range of predictor scores that are unrelated to the abilities tapped by the ASVAB.

For example, as Young et al. (1987) have described, most of the soldiers in this study enlisted in the Army between July 1983 and June 1984. They took the Project A predictor and job performance tests in the summer or fall of 1985, on average 19 months after they had reported for duty. There were many soldiers who enlisted in the Army at the same time as these soldiers who would have been eligible for our sample, but who left the Army as a result of disciplinary problems. In many instances, these problems were unrelated to any of the abilities tapped by the ASVAB. However, the problems might have been related to some of the temperaments and personality traits measured in the ABLE; indeed, several of the ABLE scales were designed to measure temperaments and traits associated with disciplinary problems. The attrition of these soldiers means that the variance of the temperament/personality scores in our soldier sample is less than the variance that we would expect to obtain in an unselected sample of 18-, 19-, and 20-year olds. Unfortunately, without data from an unselected sample, it is impossible to know the extent of this range restriction, or to correct our validity coefficients for such range restriction.

This means that many of the validity coefficients reported in the following tables are underestimates of the true validities that would be obtained in an unselected sample. The problem is probably not very serious

for the spatial ability composite or for the six perceptual-psychomotor ability composites, which are reasonably highly correlated with scores on the ASVAB. Much of the range restriction in these composites is probably alleviated by correcting for range restriction in the ASVAB. However, the problem is more serious for the composites from the three non-cognitive predictor domains. These composites tend to be relatively uncorrelated with ASVAB scores. Moreover, especially in the case of the temperament/ personality composites from the ABLE, there is reason to believe that there is a significant amount of range restriction unrelated to the abilities tapped by the ASVAB. The validities reported for these predictor domains -- and especially for the ABLE -- are likely to be underestimates of the true validities.

The procedure used to adjust R for shrinkage was developed by Claudy (1978). The adjustment is intended to yield an estimate of R that is equal to the expected value of the multiple correlation between the predictor scores and the criterion in the population from which the sample was drawn.

Relationships. Given six predictor domains and five job performance constructs, there were 30 multiple correlations generated for each of the nine jobs. (The one exception was Infantryman, which was not scored on one of the performance constructs, General Soldiering Proficiency. For Infantryman, only 24 validity coefficients were computed.) These Rs were averaged across the nine jobs to obtain the mean validity for each predictor domain by performance construct combination.

The 30 mean Rs are reported in Table 4. The table shows that the hypothesized predictor-criterion relationships (presented in Figure 7) were, by and large, confirmed.

The general cognitive ability composites, computed from the ASVAB, were the best predictors of Core Technical Proficiency (mean R = .63) and General Soldiering Proficiency (mean R = .65). These validity coefficients are extraordinarily high, especially when one considers that the ASVAB was administered to these subjects on average two years prior to the collection of job performance data. The spatial ability composite and the perceptual-psychomotor ability composites also provided excellent prediction of Core Technical Proficiency and General Soldiering Proficiency.

The general cognitive ability composites also provided reasonable prediction of Effort and Leadership (mean R = .31), as we had hypothesized it would. The mean R with Effort and Leadership was only slightly lower for the composite scores from the other two cognitive domains, spatial ability (mean R = .25) and perceptual-psychomotor ability (mean R = .26).

However, the composites within the three cognitive domains did not predict performance on Personal Discipline or Physical Fitness and Military Bearing very well. None of the six mean multiple correlations between these three predictor domains and two performance constructs exceeded .20.

The best prediction of Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing was provided by the temperament/ personality composites from the ABLE. The mean R for Effort and Leadership was .33. The ABLE composite that contributed most to this correlation was Achievement Orientation. For Personal Discipline, the mean R was .32, with the ABLE Dependability composite making the largest contribution to this R.

Table 4

Mean Validity[a] for the Composite Scores within Each Predictor Domain across Nine Army Enlisted Jo

Job Performance Construct	Predictor Domain					
	General Cognitive Ability (K=4)[b]	Spatial Ability (K=1)	Perceptual-Psychomotor Ability (K=6)	Temperament/ Personality (K=4)	Vocational Interests (K=6)	Job Reward Preferenc (K=3)
Core Technical Proficiency	.63	.56	.53	.26	.35	.29
General Soldiering Proficiency	.65	.63	.57	.25	.34	.30
Effort and Leadership	.31	.25	.26	.33	.24	.19
Personal Discipline	.16	.12	.12	.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 composites.

Finally, the ABLE composites correlated .37 on average with Physical Fitness and Military Bearing. The key predictor of this performance construct was the ABLE Physical Condition composite.

On the other hand, the temperament/personality domain provided worse prediction of the two "can do" performance criteria than any of the other five predictor domains. The mean R for Core Technical Proficiency was only .26, while the mean R for General Soldiering Proficiency was .25.

The relationships between the vocational interest composites and the job performance constructs were somewhat different than expected. For the interest composites, the pattern of correlations across the five job performance constructs was more like the pattern for the cognitive predictor domains than the pattern for the temperament/personality domain. The highest mean Rs were with Core Technical Proficiency (mean R = .35) and General Soldiering Proficiency (mean R = .34). The lowest mean Rs involved prediction of Personal Discipline (mean R = .13) and Physical Fitness and Military Bearing (mean R = .12). The mean validity for Effort and Leadership was .24.

The pattern of correlations for the job reward preference composites was similar to that for the vocational interest composites.

As a further test of the hypothesized predictor-criterion relationships presented in Figure 7, the predictor composites were grouped into two sets. The 11 general cognitive ability, spatial ability, and perceptual-psychomotor ability composites were grouped into a set of cognitive composites. The 13 temperament/personality, vocational interest, and job reward preference composites were grouped into a set of non-cognitive composites. For each set the R was computed with each of the five job performance constructs within each of the nine jobs. Mean Rs from these analyses are presented in Table 5.

The pattern of correlations is very similar to that predicted in Figure 7. The cognitive composites provide the best prediction of Core Technical Proficiency (mean R = .65) and General Soldiering Proficiency (mean R = .69). The non-cognitive composites provide the best prediction of Personal Discipline (mean R = .35) and Physical Fitness and Military Bearing (mean R = .38). The non-cognitive composites also predict Effort and Leadership better than the cognitive composites, though the difference is not very large (mean Rs = .38 and .32, respectively).

Table 5 also shows that, when all 24 composites are used to predict each performance construct, the mean Rs are .67 for Core Technical Proficiency, .70 for General Soldiering Proficiency, .44 for Effort and Leadership, .37 for Personal Discipline, and .42 for Physical Fitness and Military Bearing. These results indicate that for at least two of the job performance constructs -- Effort and Leadership and Physical Fitness and Military Bearing -- the best prediction is obtained when both cognitive and non-cognitive predictors are used.

The one surprising result in Table 5 is the high correlation between the non-cognitive predictors and the two "can do" performance constructs. For both performance constructs, the mean R was .44. In fact, the non-cognitive composites predicted "can do" performance better than they predicted "will do" performance.

Table 5

Mean Validity[a] for the Cognitive, the Non-Cognitive, and All Predictor Composites across Nine Army Enlisted Jobs

Job Performance Construct	Predictor Composites		
	Cognitive (K=11)[b]	Non-Cognitive (K=13)	All (K=24)
Core Technical Proficiency	.65	.44	.67
General Soldiering Proficiency	.69	.44	.70
Effort and Leadership	.32	.38	.44
Personal Discipline	.17	.35	.37
Physical Fitness and Military Bearing	.23	.38	.42

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b]K is the number of predictor composites.

The Incremental Validity of the Project A Predictor Tests

An important question for the Army sponsors of the present study was 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 composite scores (computed from the ASVAB) was compared to the validity obtained when the composite scores from a predictor domain were used to supplement the general cognitive ability composites. 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 6.

Table 6 shows that none of the predictor domains added more than .02 to the general cognitive ability composites' validity for predicting Core Technical Proficiency. Similarly, no predictor domain added more than .03 to the general cognitive ability composites' validity for predicting General Soldiering Proficiency. In both instances, the predictor composite that added the greatest incremental validity was the spatial ability composite.

Most predictor domains also added very little to the prediction of Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing from the general cognitive ability composites. The one exception was the temperament/personality domain. The four temperament/personality composites added .11 to the validity for predicting Effort and Leadership, .19 to the validity for predicting Personal Discipline, and .21 to the validity for predicting Physical Fitness and Military Bearing.

Table 7 provides another means for looking at the incremental validity of the Project A predictor tests. The table shows that the seven new Project A cognitive composites (i.e., the spatial ability composite plus the six perceptual-psychomotor ability composites) predict job performance almost as well as the four general cognitive ability composites from the ASVAB. For Core Technical Proficiency and General Soldiering Proficiency, the validity of the new Project A cognitive composites is quite high (mean $\underline{R}$ = .59 and .65, respectively). However, the performance variance predicted by the new Project A cognitive composites is virtually identical to the performance variance predicted by the ASVAB. The new Project A cognitive composites increment the validity for Core Technical Proficiency by .02 and increment the validity for General Soldiering Proficiency by .04. (At first glance, those results were disappointing to many of us on the Project A research team. However, as we had time to reflect, we decided that we had established that the Army was already doing a very good job of predicting "can do" job performance, which our Army sponsors were pleased to hear. Also, as a practical matter, there simply isn't much that one can do to improve on a test with a validity of .63 or .65 for predicting job performance two years later.)

Table 7 also shows that the 13 non-cognitive composites predict Effort and Leadership (mean $\underline{R}$ = .38), Personal Discipline (mean $\underline{R}$ = .35), and Physical Fitness and Military Bearing (mean $\underline{R}$ = .38) better than the four general cognitive ability composites predict these three job performance constructs. When the ASVAB composites are added to the non-cognitive

Mean Incremental Validity[a,b] for the Composite Scores within Each Predictor Domain across Nine Army Enlisted Jobs

Job Performance Construct	Predictor Domain: General Cognitive Ability (K=4)[c]	General Cognitive Ability Plus Spatial Ability (K=5)	General Cognitive Ability Plus Perceptual-Psychomotor Ability (K=10)	General Cognitive Ability Plus Temperament/ Personality (K=8)	General Cognitive Ability Plus Vocational Interests (K=10)	General Cognitive Ability Plus Job Reward Preference (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.
[b]Incremental validity refers to the increase in R afforded by the new predictors above and beyond the R for the Army's current predictor battery, the ASVAB.
[c]K is the number of predictor composites.

Table 7

Mean Validity[a,b] for the Project A Cognitive and the Project A Non-Cognitive Predictor Composites across Nine Army Enlisted Jobs

Job Performance Construct	Predictor Composites				
		Cognitive		Non-Cognitive	
	General Cognitive Ability (ASVAB) Composites (K=4)[c]	New Project A Cognitive Composites (K=7)	New Project A Cognitive Composites Plus ASVAB Composites (K=11)	New Project A Non-Cognitive Composites (K=13)	New Project A Cognitive Composites Plus ASVAB Composites (K=17)
Core Technical Proficiency	.63	.59	.65	.44	.65
General Soldiering Proficiency	.65	.65	.69	.44	.67
Effort and Leadership	.31	.27	.32	.38	.43
Personal Discipline	.16	.13	.17	.35	.37
Physical Fitness and Military Bearing	.20	.14	.23	.38	.41

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b]Incremental validity refers to the increase in R afforded by the new predictors above and beyond the R for the Army's current predictor battery, the ASVAB.
[c]K is the number of predictor composites.

composites, the mean validity for Effort and Leadership increases by .05, the mean validity for Personal Discipline increases by .02, and the validity for Physical Fitness and Military Bearing increases by .03.

The results in Table 7 are consistent with our hypotheses (see Figure 7) that: (1) cognitive ability composites would be the most valid predictors of Core Technical Proficiency and General Soldiering Proficiency; (2) non-cognitive composites would be the most valid predictors of Personal Discipline and Physical Fitness and Military Bearing; and (3) both cognitive and non-cognitive predictors would be useful for predicting Effort and Leadership.

A comparison of Tables 6 and 7 shows that almost all of the incremental validity in the prediction of the three "will do" performance constructs is provided by the ABLE. When the ABLE composites and the ASVAB composites are used to predict Effort and Leadership the mean R is .42. When the AVOICE composites and the JOB composites are added to the ABLE and ASVAB composites, the mean validity increases only by .01. Similarly, the AVOICE and JOB composites add only .02 to the prediction of Personal Discipline and contribute nothing to the prediction of Physical Fitness and Military Bearing.

Relationships between Predictor Domains and "Method Factors"

In their paper, Campbell et al. (1987) described written test and rating "method factors" that emerged from a structural analysis of the job performance measures. As Campbell et al. noted, the term "method factor" is probably a misnomer. It is likely that these factors represent important components of job performance.

The written test factor reflects a soldier's comprehension of the manuals, instructions, and other materials that must be read on the job. For several of the jobs that were studied, excerpts from technical manuals and other learning aids were incorporated into the written knowledge tests. It is likely that a soldier who had difficulty reading and comprehending these materials during Project A performance testing also would have difficulty using these written materials on the job.

The rating factor represents raters' global impressions of soldiers. It is similar to what many researchers might term "halo error" (cf. reference, 19xx). There is, however, no proof that this rating factor truly is error. It is equally possible that the global impression represented by the rating factor is an important measure of soldier effectiveness. The Project A data base provides an opportunity to study the relationships between this rating factor and individual difference variables from several domains.

Table 8 shows the multiple correlations between the predictors within each domain and the two method factors. The mean Rs for the written test factor are much greater than the mean Rs for the rating factor across all six predictor domains.

The best predictors of the written test factor were the general

Table 8

Mean Validity[a] for the Composite Scores within Each Predictor Domain across Nine Army Enlisted Jobs for Written Test and Rating "Method Factor" Scores

Method Factor	Predictor Domain					
	General Cognitive Ability (K=4)[b]	Spatial Ability (K=1)	Perceptual-Psychomotor Ability (K=6)	Temperament/ Personality (K=4)	Vocational Interests (K=6)	Job Reward Preferences (K=3)
Written Test	.62	.55	.54	.21	.32	.28
Rating	.15	.07	.08	.18	.09	.08

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b]K is the number of predictor composites.

cognitive ability composites (mean $\underline{R}$ = .62). Across the nine jobs the ASVAB verbal composite was the most consistent predictor of the written test factor. The spatial ability composite and the perceptual-psychomotor ability composites had mean correlations of .55 and .54, respectively. Correlations were lower for the composites within the three non-cognitive domains. However, the mean correlations were not trivial, ranging from .21 for the temperament/personality composites to .32 for the vocational interest composites. This pattern of correlations contributes additional evidence that this factor represents a soldier's proficiency at reading job-related materials.

The best predictors of the rating factor were the temperament/personality composites (mean $\underline{R}$ = .18). Within the temperament/personality domain, the most consistent predictor of the rating factor was the ABLE dependability composite. After the temperament/personality composites, the second best predictors were the general cognitive ability composites (mean $\underline{R}$ = .15). Mean correlations for the composites within the remaining four domains all were less than .10. This pattern of correlations suggests that the rating factor taps dependability on the job, but much more evidence would be needed to confirm this interpretation.

For Table 9, the predictor composites again were grouped into two sets. For the written test factor, the mean $\underline{R}$s across the nine jobs were .64 for the 11 cognitive composites, .40 for the 13 non-cognitive composites, and .65 across all 24 predictor composites. For the rating factor, the mean $\underline{R}$s were .16, .22, and .26, respectively.

The pattern of correlations for the rating factor is similar to the pattern for the Effort and Leadership performance construct (see Table 5). This suggests that the rating factor obtained in this study reflects raters' global impressions of soldiers' overall competency and dependability. That is, when raters were asked to evaluate a soldier on a particular rating dimension, they considered the soldier's performance on that dimension and two other factors as well. The first factor was their general impression of how well the soldier was capable of performing the job. The second was their general impression of the soldier's dependability.

Another method of studying the two method factors is to examine how the pattern of predictor-criterion relationships changes when the variance attributable to the method factors is removed from the five performance construct scores. These results are presented in Table 10.

The validity coefficients presented for the "raw" performance construct scores in Table 10 are the same as those presented in Table 4. To compute residual performance construct scores, the variance attributable to the written test factor was partialed from the scores for Core Technical Proficiency and General Soldiering Proficiency, and the variance attributable to the rating factor was partialed from the scores for Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing. (Written knowledge tests were not used in computing scores for Effort and Leadership, Personal Discipline, or Physical Fitness and Military Bearing. Nor were rating scales used in computing scores for Core Technical Proficiency or General Soldiering Proficiency.)

The table shows that the residual scores for Core Technical Proficiency and General Soldiering Proficiency were much less predictable than the raw

Table 9

Mean Validity[a] for the Cognitive, the Non-Cognitive, and All Predictor Composites across Nine Army Enlisted Jobs for Written Test and Rating "Method Factor" Scores

	Predictor Composites		
Method Factor	Cognitive (K=11)[b]	Non-Cognitive (K=13)	All (K=24)
Written Test	.64	.40	.65
Rating	.16	.22	.26

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b]K is the number of predictor composites.

Table 16

Mean Validity [a] for the Composite Scores within Each Predictor Domain across Nine Army Enlisted Jobs for "Raw" and "Residual" Job Performance Construct Scores

Job Performance Construct	Type of Score	Predictor Domain: General Cognitive Ability (K=4) [b]	Spatial Ability (K=1)	Perceptual-Psychomotor Ability (K=6)	Temperament/ Personality (K=4)	Vocational Interests (K=6)	Job Reward Preferences (K=3)
Core Technical Proficiency	Raw	.63	.56	.53	.26	.35	.29
	Residual	.47	.37	.37	.22	.28	.21
General Soldiering Proficiency	Raw	.65	.63	.57	.25	.34	.30
	Residual	.49	.48	.41	.21	.26	.22
Effort and Leadership	Raw	.31	.25	.26	.33	.24	.19
	Residual	.46	.41	.38	.31	.32	.27
Personal Discipline	Raw	.16	.12	.12	.32	.13	.11
	Residual	.19	.15	.13	.28	.15	.10
Physical Fitness and Military Bearing	Raw	.20	.10	.11	.37	.12	.11
	Residual	.21	.11	.14	.35	.14	.10

[a] Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b] K is the number of predictor composites.

scores. This was true across all six predictor domains. The decrease in the mean R was greater for the cognitive predictor domains than for the non-cognitive predictor domains.

For Effort and Leadership, the cognitive predictor cognitive predicted the residual performance construct scores better than they predicted the raw performance construct scores. For example, the mean R of the general cognitive ability composites with the raw Effort and Leadership score was .31, while the mean R with the residual Effort and Leadership score was .46. Thus, the mean R was .15 higher for the residual score than for the raw score. The increase was .16 for the spatial ability composite (mean R = .41 for residual Effort and Leadership and .25 for raw Effort and Leadership) and .12 for the perceptual-psychomotor ability composites (mean R = .38 and .26 for residual and raw Effort and Leadership scores, respectively).

For the temperament/personality composites, the results were the opposite. The mean multiple correlation of the temperament/personality composites with the raw Effort and Leadership score was .33, while the mean R with the residual score was .31.

The vocational interest composites and the job reward preference composites actually "behaved" similarly to the cognitive ability composites. For both predictor domains, the mean R was greater for the residual Effort and Leadership score than for the raw Effort and Leadership score.

This pattern of correlations for Effort and Leadership suggests two interesting conclusions. First, the pattern provides additional evidence that the vocational interest composites are more similar to cognitive predictors than to temperament/personality predictors.

Second, the changes in the pattern of correlations between raw and residual scores suggest that Effort and Leadership becomes more like a "can do" performance construct when the rating method factor is partialed from the raw score. The mean multiple correlations between the residual Effort and Leadership score and the cognitive predictor composites are very similar to the mean Rs between the two residual proficiency construct scores and the cognitive predictor composites. On the other hand, the residual Effort and Leadership score has a much higher correlation with the temperament/personality composites than the two residual proficiency construct scores have (mean R = .31 for Effort and Leadership, .22 for Core Technical Proficiency, and .21 for General Soldiering Proficiency). This indicates that, even after the rating factor is partialed from the raw Effort and Leadership score, the residual Effort and Leadership score continues to reflect the "will do" portion of the job performance space. Thus, the residual Effort and Leadership score appears to tap both "can do" or maximal job performance and "will do" or typical job performance.

Partialing the rating factor from the Personal Discipline and Physical Fitness and Military Bearing scores had little impact on the correlations of these scores with the predictor composites. None of the correlations for these two performance constructs changed by more than .04 when residual scores were used instead of raw scores.

Summary and Conclusions

The pattern of predictor-criterion relationships presented in this paper was consistent with the pattern that was expected. Cognitive predictors provided excellent prediction of Core Technical Proficiency and General Soldiering Proficiency. Across nine very different jobs, the mean R for the complete set of 11 cognitive composite scores was .65 for Core Technical Proficiency and .69 for General Soldiering Proficiency. Clearly, cognitive predictors provide excellent prediction of job proficiency for Army enlistees. Non-cognitive predictors -- specifically, temperament/ personality predictors -- were the best predictors of Personal Discipline and Physical Fitness and Military Bearing. The best prediction of Effort and Leadership was obtained when both cognitive and non-cognitive predictors were used.

The predictor-criterion relationships uncovered enhanced understanding of both the predictor space and the job performance space. On the predictor side, the vocational interest composites provided surprisingly good prediction of Core Technical Proficiency and General Soldiering Proficiency. In retrospect, these correlations often made perfectly good sense. For example, as Wise, Campbell, and Peterson (1987) note, the combat-related interest composite was correlated with scores on General Soldiering Proficiency, which represents performance on common soldiering tasks. The combat-related interest composite also was correlated with Core Technical Proficiency scores in the the three combat jobs studied (Infantryman, Cannon Crewmember, and Armor Crewman). In retrospect, these correlations often made perfectly good sense -- as research results often do, in retrospect. In this case, the results suggest that people who are interested in their work are more likely to perform well on their job than people who are not interested in their work. This certainly is not surprising, in retrospect.

On the criterion side, the pattern of predictor-criterion correlations helped add to our confidence in the construct validity of the job performance scores. The pattern of correlations also enhanced understanding of the Effort and Leadership construct, the written test and rating method factors, and the relationship between raw and residual performance construct scores.

The correlations of the vocational interest and job reward preference composites with the "will do" performance criteria point to one weakness of the Project A criterion measures. The best criteria for these predictors would be some measure of job satisfaction. In future Project A validation research, we will include job satisfaction measures in our assessment.

References

Campbell, J. P., Harris, J. H., McHenry, J. J., & Arabian, J. (1987, April). Analysis of criterion measures: The modeling of performance. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Claudy, J. G. (1978). Multiple regression and validity estimation in one

sample. Applied Psychological Measurement, 2, 4, 595-601.

Lord, P., & Novick, M. (1968). Statistical theory of mental test scores. Reading, MA: Addison-Wesley Publishing Company, Inc.

Mitchell, K. J., Hanser, L. M., & Grafton, F. C. (1984). The 1980 youth population norms: enlisted and occupational classification standards in the Army (ARI-RS-WP-84-13). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. A., Houston, J. S., Toquam, J. L., & Wing, H. (1987, April). Identification of predictor constructs and development of new selection/classification tests. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Young, Y. Y., Harris, J. H., Hoffman, G. R., & Houston, J. S. (1987, April). Large scale data collection and data base preparation. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Wing, H., Peterson, N. G., & Hoffman, R. G. (1984, August). Expert judgments of predictor-criterion validity relationships. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Canada.

Wise, L. L., Campbell, J. C., & Peterson, N. G. (1987, April). Identifying optimal predictor composites for generalizability across jobs and performance constructs. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

IDENTIFICATION OF PREDICTOR CONSTRUCTS AND DEVELOPMENT OF NEW SELECTION/CLASSIFICATION TESTS

Norman G. Peterson
Leaetta M. Hough
Marvin D. Dunnette
Rodney L. Rosse
Janis S. Houston
Jody L. Toquam
Personnel Decisions Research Institute

Hilda Wing
U.S. Army Research Institute

Presented at the Annual Conference of the
Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

The distinguished, earlier speakers have elegantly and accurately described the scope and purpose of Project A. Before plunging into the main topic of this paper, however, we would like to present a few figures that show the real questions that had to be addressed by Project A.

Figure 1 shows, at first blush, what appeared to be the major question to be answered. Well, I have to admit, this question had some of us a little bit worried.

Imagine our relief then, when we discovered that the real question to be answered was the one shown in Figure 2.

Finally, Figure 3 shows the question posed for the predictor team of Project A. Well, by now we were down to a question that any reality-grounded psychologist could really be afraid of tackling.

Anyway, that was where we began. In the remainder of this paper, we present an overview of the process followed to address the question in Figure 3, and the battery of tests developed through that process.

APPROACH AND RESEARCH DESIGN

Theoretical Approach

At present, the U.S. Army has a large number of jobs (called Military Occupational Specialities or MOS) and hires, almost exclusively, inexperienced and untrained persons to fill those jobs. As obvious as these facts are, they need to be stated because they are the overriding facts that have to be addressed by the predictor team on Project A.

One implication of these facts is that a highly varied set of individual differences' variables must be put into use if there is to be a reasonable chance of improving the present level of accuracy of predicting training performance, job performance, and attrition/retention in a substantial proportion, if not all, of those jobs. Much less evident is the particular content of that set of individual differences variables, and the way the set should be developed and organized.

A second, and perhaps less obvious, implication is the notion that new predictor measures must be appropriate for selecting persons who do not have the training and experience to begin immediately performing their assigned jobs. This is true partly because of the vast numbers of job positions that need to be filled, partly because of the kinds of jobs found in the Army (Infantry, Artillery, etc.), and partly because of the population of persons that the Army draws from (young high-school graduates with little or no specialized training and job experience).

These considerations led us to adopt a construct-oriented strategy of predictor development, but with a healthy leavening from the content-oriented strategy. Essentially, we endeavored to build up a model of predictor space by (1) identifying the major, relatively independent domains or types of

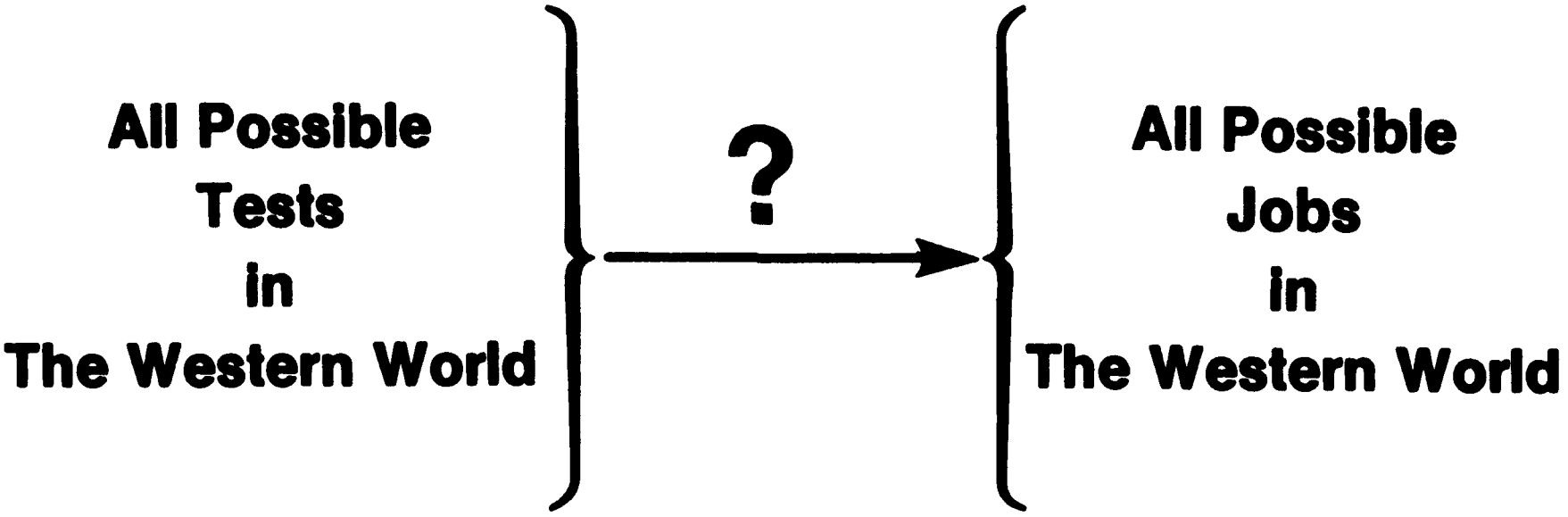


Figure 1. Project A: Question To be Answered

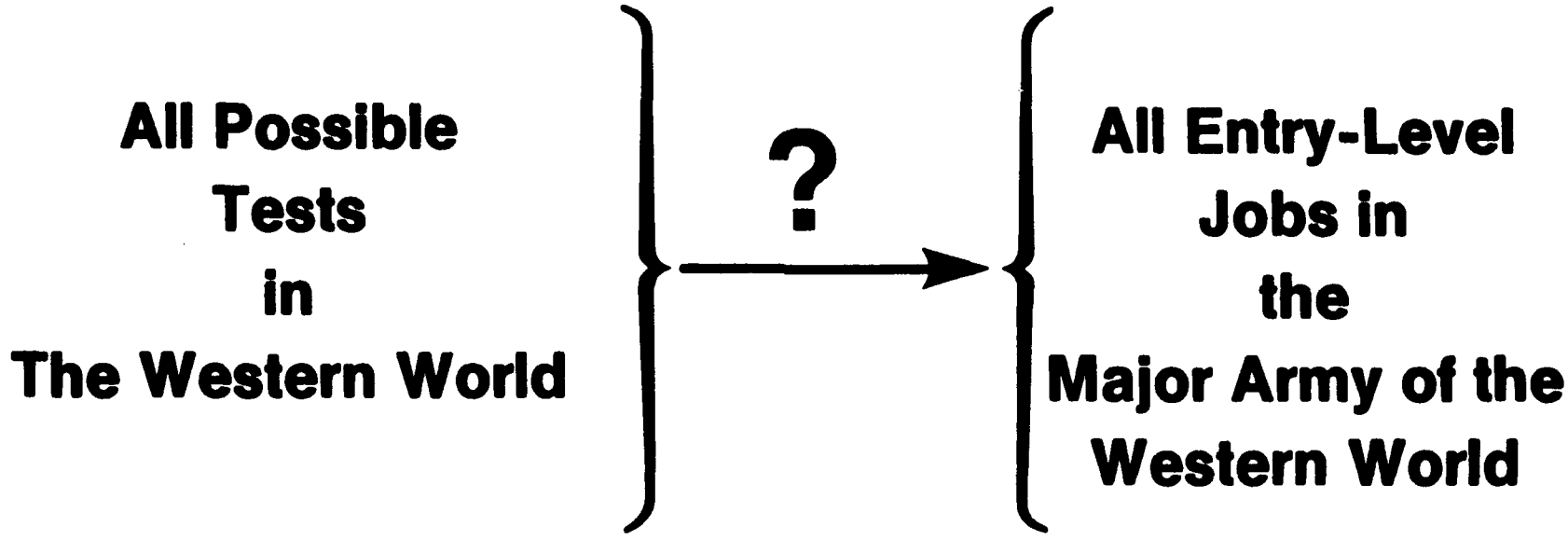


Figure 2. Project A: Question To be Answered

What Predictor Constructs Will Add the Most Validity For Predicting Total Job Performance and Increase the Accuracy of Job Placement?

P.S. (Once these constructs are found, go ahead and develop the best possible tests for them, administer them to thousands of soldiers, and report the incremental validities.)

Figure 3. Project A: Question to be Answered By the Project A Predictor Team.

individual differences' constructs that existed; (2) selecting measures of constructs within each domain that met a number of psychometric and pragmatic criteria; and, (3) further selecting those constructs that appeared to be the "best bets" for incrementing (over present predictors) the prediction of the set of criteria of concern (i.e., training/job performance and attrition/ retention in Army jobs).

Ideally, the model would, we hoped, lead to the selection of a finite set of relatively independent predictor constructs that were also relatively independent of present predictors and maximally related to the criteria of interest. If these conditions were met, then the resulting set of measures would predict all or most of the criteria, yet possess enough heterogeneity to yield powerful, efficient classification of persons into different jobs.

The development of such a model also had the virtue that it could be at least partially "tested" at many points during the research effort, and not just at the end, when all the predictor and criterion data are in. For example, we could examine the covariance of newly developed measures with one another and with the present predictors, notably the Armed Services Vocational Aptitude Battery (ASVAB). If the new measures were not relatively independent of the ASVAB and measures from other domains as predicted by the model, then we could take steps to correct that. Also, by constructing such a visible model, we thought that modifications and improvements could be implemented much more straightforwardly.

Figure 4 shows an illustrative, construct-oriented model and is presented in order to represent the model in abstract. Note that both the criterion and the predictor space are depicted. As mentioned earlier, a great deal of the work of Project A is devoted to the development of criterion measures, and we, on the predictor side, have taken advantage of the information coming from those efforts as it has become available.

If this illustrative model were to be developed and tested with data, then the network of relationships on the predictor side, on the criterion side, and between the two could be confirmed, disconfirmed, and/or modified. It is imperative that the development of such models be done very carefully and conservatively, and subjected frequently to reality testing; we have kept this firmly in mind. However, the possession of such a model enables one to state fairly clearly why such and such a predictor is being researched, and to check quickly, at least rationally, whether the addition of a predictor is likely to improve prediction.

Finally, the model is depicted as a matrix with a hierarchical arrangement of both the rows and the columns. We have found it useful to employ this hierarchical notion, because it allows us to think in terms of appropriate levels of specificity for a particular problem as we do the research, or for future applications of measures.

Research Objectives - Destinations Along the Way

This theoretical approach led to the delineation of seven more concrete objectives of our research. These were:

1. Identify measures of human abilities, attributes, or characteristics which are most likely to be effective in predicting, prior to

PREDICTORS		CRITERIA							
		Training Performance			Job Task Performance		Attrition/ Retention		
		Pass/ Fail	Test Grades	Atten-dance	Common Tasks	Specific Tasks	Finish Term	Reen-list	Early Discharge
Cognitive	Verbal	M*	H	L	M	M	L	L	L
	Numerical	M	H	.	.	.	.		
	Spatial								
Psychomotor	Precision								
	Coordination								
	Dexterity								
Temperament	Dependability								
	Dominance								
	Sociability								
Interests	Realistic								
	Artistic								
	Social	.	.	.	M	M	M	L	L

*Denotes expected strength of relationship, High, Medium, Low.

Figure 4. Illustrative construct-oriented model.

entry into the organization, successful performance in general, and in classifying persons into jobs where they will be most successful, with special emphasis on attributes not tapped by current preinduction measures.

2. Design and develop new measures or modify existing measures of these "best bet" predictors.

3. Develop materials and procedures for efficiently administering experimental predictor measures in the field.

4. Estimate and evaluate the reliability of the new preinduction measures and their vulnerability to motivational set differences, faking, variances in administrative settings, and practice effects.

5. Determine the interrelationships (or covariance) between the new preinduction measures and current preinduction measures.

6. Determine the degree to which the validity of new preinduction measures generalizes across jobs; that is, proves useful for predicting measures of successful performance across quite different jobs and, conversely, the degree to which the measures are useful for classification or the differential prediction of success across jobs.

7. Determine the extent to which new preinduction measures increase the accuracy of prediction of success and the accuracy of classification into jobs over and above the levels of accuracy reached by current preinduction measures.

Research Design - The Road Map

To achieve these objectives, we have followed the design depicted in Figure 5.

Several things, we feel, are noteworthy about the design. First, five test batteries are mentioned: Preliminary Battery, Demo Computer Battery, Pilot Trial Battery, Trial Battery, and Experimental Battery. These appear successively in time and allow us to modify and improve our predictors as we gather and analyze data on each successive battery or set of measures.

Second, a large-scale literature review and a quantified expert judgment process were used early in the project to take maximum advantage of earlier research and accumulated knowledge and expert opinion. The expert judgment process was used to develop an early model of both the predictor space and the criterion space and relied heavily on the information gained from the literature review. By using the model that resulted from analyses of the experts' judgments of the relationships between predictor constructs and criterion dimensions, we were able to develop, carefully and efficiently, measures of the most promising predictor constructs.

Third, the design includes both predictive (for the Preliminary and Experimental Batteries) and concurrent (for the Trial Battery) validation modes of data collection, although that is not obvious from Figure 5. Thus,

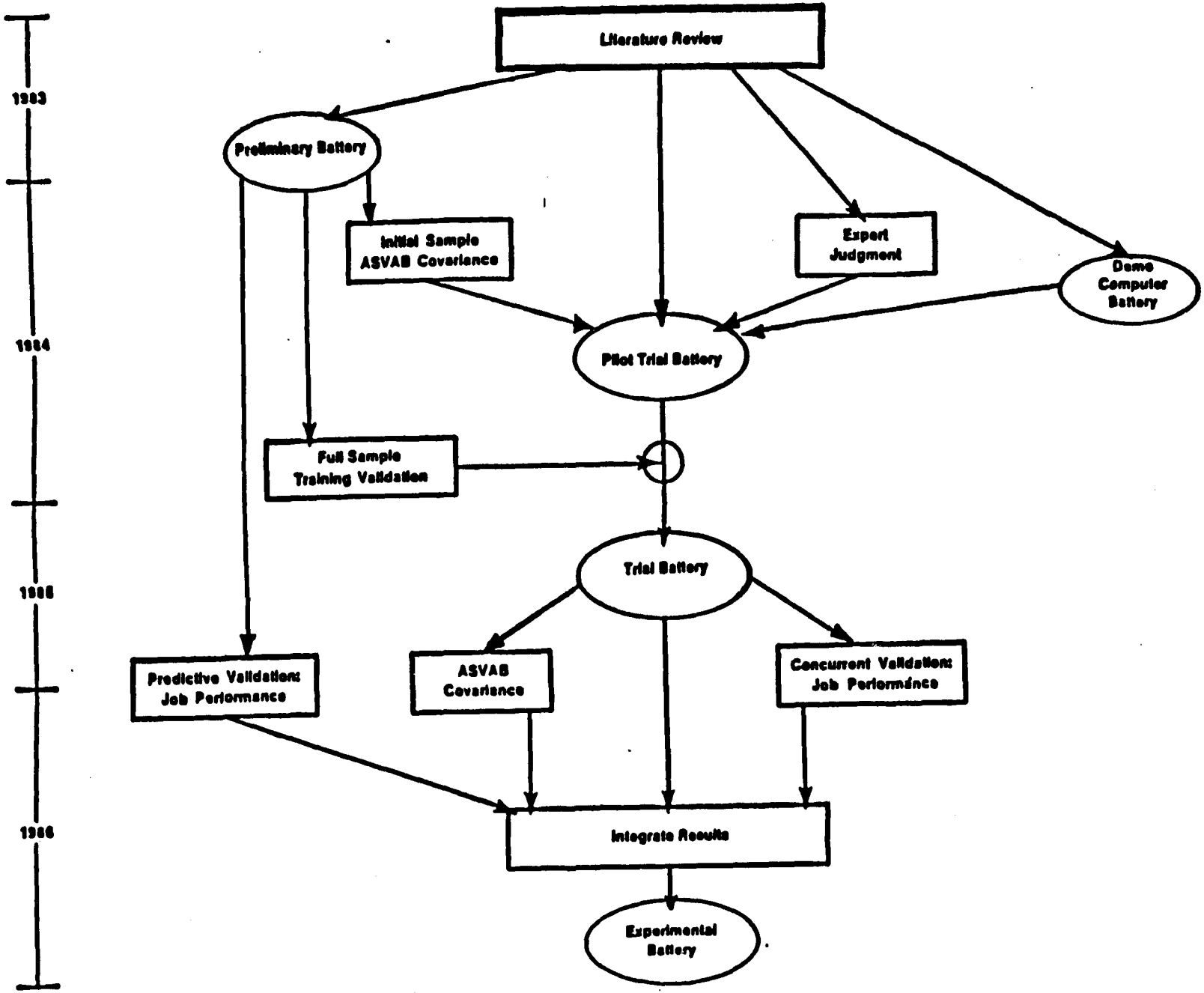


Figure 5. Flow chart of predictor measure development activities of Project A.

we are able to benefit from the advantage of both types of designs, -- that is, early collection and analysis of empirical criterion-related validities in the case of the concurrent design, and less concern about range restriction and experiential effects in the predictive design.

Organization

We organized predictor researchers into three "domain teams" as we worked our way through this research design and toward the earlier described research objectives. One team concerned itself with the temperament, biographical data, and vocational interest variables and came to be called the "non-cognitive" team. Another team concerned itself with cognitive and perceptual kinds of variables and was called the "cognitive" team. The third team concerned itself with psychomotor and perceptual variables and was labeled the "psychomotor" team or sometimes the "computerized" team since all the measures developed by that team were computer-administered.

Another important component in the organization was the set of scientific advisors assigned to overlook and assist us, particularly, Lloyd Humphreys and Jay Uhlaner. These scientists met frequently with, and advised us at critical decision points. The experience and wisdom they brought to the team were extremely valuable.

INITIAL RESEARCH STEPS

The overriding purpose of the literature review was, simply put, to make maximum use of earlier research on the problem of accurately predicting job performance and classifying persons into jobs in such a way that both the person and the organization receive maximum benefits. More specifically, we wished to identify those variables or constructs, and their measures, that had proven effective for such purposes. As Figure 5 shows, the information obtained from the literature review was used in all the immediately succeeding research activities.

The search was conducted by the three research teams, each responsible for a fairly broadly defined area of human abilities or characteristics: cognitive abilities; non-cognitive characteristics such as vocational interests, biographical data, and measures of temperament; and psychomotor/physical abilities.

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

1. Compile an exhaustive list of possibly relevant reports, articles, books, or other sources using computerized data base searches, existing bibliographies, and consultation of experts.

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

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

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

Across all three ability areas, more than 10,000 sources were identified via the computer search. (Of course, many of these sources were identified as relevant in more than one area, and were thus counted more than once.)

The special review forms and the actual sources that had been located were used in two primary ways. First, three working documents were written, one for each of the three areas. (These documents were put into research note form: Hough, Kamp & Barge, in press; Toquam, Corpe, & Dunnette, in press; McHenry & Rose, in press.) These documents identified and summarized the literature with regard to issues important to the research being conducted, the most appropriate organization or taxonomy of the constructs in each area, and the validities of the various measures for different types of jobs performance criteria. Second, the predictors identified in the review were subjected to further, structured scrutiny in order to (1) select tests and inventories to make up the Preliminary Battery, and (2) select the "best bet" predictor constructs to be used in the expert judgment research activity.

Expert Judgments

The approach used in the expert judgment process was to (1) identify criterion categories, (2) identify an exhaustive range of psychological constructs that may be potentially valid predictors of those criterion categories, and (3) obtain expert judgments about the relationships between the two. Schmidt, Hunter, Croll, and McKenzie (1983) showed that pooled expert judgments, obtained from experienced personnel psychologists, were as accurate in estimating the validity of tests as actual, empirical criterion-related validity research using samples of hundreds of subjects. That is, experienced personnel psychologists are effective "validity generalizers" for cognitive tests. They do tend to underestimate slightly the true validity as obtained from empirical research.

Hence, one way to identify the "best best" set of predictor variables and measures is to use a formal judgment process employing experts such as that followed by Schmidt et al., (1983). Descriptive information about a set of predictors and the job performance criterion variables is given to "experts" in personnel selection and classification, typically personnel psychologists. These experts estimate the relationships between predictor and criterion variables by rating or directly estimating the value of the correlation coefficients.

The result is a matrix with predictor and criterion variables as the columns and rows, respectively. Cell entries are experts' estimates of the degree of relationship between the particular predictors and various criteria. The interrater reliability of the experts' estimates is checked first. If the estimate is sufficiently reliable (previous research shows values in the .80 to .90 range for about 10 to 12 experts), the matrix of predictor-criterion relationships can be analyzed and used in a variety of ways. By correlating the columns of the matrix, the covariances of the

predictors can be estimated on the basis of the profiles of their estimated relationships with the criteria. These covariances can then be factor analyzed to identify predictors that function similarly in predicting performance criteria. Similarly, the criterion covariances can be examined to identify clusters of criteria predicted by a common set of predictors.

Such procedures help identify redundancies and overlap in the predictor set. The common sets or clusters of predictors and of criteria are an important product for several reasons. Most importantly here, these clusters provide a model or theory of predictor-criterion performance space. This model serves as an informative guide to development of a set of predictors that should be efficient and valid, at least insofar as the informed opinion of knowledgeable experts can propel one in that direction.

To carry out the expert judgment activity, we had to identify predictor and criterion variables and prepare materials that would enable the experts to provide reliable estimates of validity. Time does not permit a description of these activities. The predictor team identified the predictor variables, while the Project A criterion team(s) identified the criterion variables.

In the end, we had 35 experts rate the validity of 53 predictor variables for 72 criterion variables,, using materials and instructions prepared by us. Results showed that the means of the predictor-criterion validity judgments (cell means) were highly reliable (.96), and factor analysis revealed eight predictor factors that summarized the judgments of the experts. Scrutiny of these findings resulted in the hierarchical model shown in Figure 6.-

The expert judgment task then, resulted in a hierarchical model of predictor space that served as a guide for the development of new, preinduction measures (the Pilot Trial Battery, See Figure 5) for Army enlisted ranks. (Wing, Peterson, and Hoffman, 1984, provide a detailed presentation of the expert judgment process and results.)

Preliminary Battery

The Preliminary Battery (PB) was conceived of as a set of proven "off-the shelf" measures of predictors that overlapped very little with the Army's current pre-induction predictors. The collection of data on a number of predictors that represent the types of predictors not currently in use by the Army would allow an early determination of the extent to which such predictors contributed unique variance, that is, measured attributes not measured by current pre-induction predictors. This information would be useful for guiding the development of new predictors into areas most likely to be useful for increasing the accuracy of prediction and classification.

Also, the collection of predictor data (from soldiers in training) early in the project allowed the conduct of a predictive validity investigation much earlier in the project than if we were to wait until the Trial Battery was developed.

CONSTRUCTS	CLUSTERS	FACTORS
1. Verbal Comprehension 5. Reading Comprehension 16. Ideational Fluency 18. Analogical Reasoning 21. Omnibus Intelligence/Aptitude 22. Word Fluency	A. Verbal Ability/ General Intelligence	
4. Word Problems 8. Inductive Reasoning: Concept Formation 10. Deductive Logic	B. Reasoning	
2. Numerical Computation 3. Use of Formula/Number Problems	C. Number Ability	COGNITIVE ABILITIES
12. Perceptual Speed and Accuracy	H. Perceptual Speed and Accuracy	
49. Investigative Interests	N. Investigative Interests	
14. Rote Memory 17. Follow Directions	J. Memory	
19. Figural Reasoning 23. Verbal and Figural Closure	F. Closure	
6. Two-dimensional Mental Rotation 7. Three-dimensional Mental Rotation 9. Spatial Visualization 11. Field Dependence (Negative) 15. Place Memory (Visual Memory) 20. Spatial Scanning	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
24. Processing Efficiency 25. Selective Attention 26. Time Sharing	G. Mental Information Processing	INFORMATION PROCESSING
13. Mechanical Comprehension	L. Mechanical Comprehension	MECHANICAL
48. Realistic Interests 51. Artistic Interests (Negative)	M. Realistic vs. Artistic Interests	
28. Control Precision 29. Rate Control 32. Arm-hand Steadiness 34. Aiming	I. Steadiness/Precision	
27. Multilimb Coordination 35. Speed of Arm Movement	D. Coordination	PSYCHOMOTOR
30. Manual Dexterity 31. Finger Dexterity 33. Wrist-Finger Speed	K. Dexterity	
39. Sociability 52. Social Interests	Q. Sociability	SOCIAL SKILLS
50. Enterprising Interests	R. Enterprising Interests	
36. Involvement in Athletics and Physical Conditioning 37. Energy Level	T. Athletic Abilities/Energy	VIGOR
41. Dominance 42. Self-esteem	S. Dominance/Self-esteem	
40. Traditional Values 43. Conscientiousness 46. Non-delinquency 53. Conventional Interests	N. Traditional Values/Conventionality/Non-delinquency	
44. Locus of Control 47. Work Orientation	O. Work Orientation/Locus of Control	MOTIVATION/ STABILITY
38. Cooperativeness 45. Emotional Stability	P. Cooperation/Emotional Stability	

Figure. 6. Hierarchical map of predictor space.

Selection of Preliminary Battery Measures

The literature review identified a large set of predictor measures, each with ratings by the researchers on several psychometric and substantive evaluation factors. These ratings wee used to select a smaller set of measures as serious candidates for inclusion in the Preliminary Battery. Two major practical constraints came into play: (1) no apparatus or individualized testing methods could be used because of the relatively short time available to prepare for battery administration, and the fact that the battery would be administered to a large number of soldiers (several thousand) over a nine-month period by relatively unsophisticated test administrators, and (2) only four hours were available for testing.

Predictor team researchers, and several prominent scientists outside the predictor team, made the selection of "off-the-shelf" measures.

The Preliminary Battery included the following:

- Eight perceptual-cognitive measures
  - Five from the Educational Testing Service (ETS) French Kit (Ekstrom, French, and Harman, 1976)
  - Two from the Employee Aptitude Survey (EAS) (Ruch and Ruch, 1980)
  - One from the Flanagan Industrial Tests (FIT) (Flanagan, 1965)
- Eighteen scales from the Air Force Vocational Interest Career Examination (VOICE) (Alley and Matthews, 1982).
- Five temperament scales adapted from published scales
  - Two from the Differential Personality Questionnaire (DPQ)
  - One from the California Psychological Inventory (CPI) (Gough, 1975)
  - The Rotter I/E scale (Rotter, 1966)
  - Validity scales from both the DPQ and the Personality Research Form (PRF) (Jackson, 1967)
- Owen's Biographical Questionnaire (BQ) (Owens and Schoenfeldt, 1979). The BQ could be scored for either 11 scales for males or 14 for females, based on Owen's research, or for 18 predesignated, combined-sex scales developed for this research and called Rational Scales. The rational scales had no item on more than one scale; some of Owen's scales included items on more than one scale. Items tapping religious or socio-economic status were deleted from Owens' instrument for this use, and items tapping physical fitness and vocational-technical course work were added.

In addition to the Preliminary Battery, scores were available for the Armed Services Vocational Aptitude Battery, which all soldiers take prior to entry into service.

Sample and Administration of Battery

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

Analyses

An initial set of analyses was performed on the Preliminary Battery data to inform the development of the Pilot Trial Battery (PTB). (The PTB was intended to include newly developed tests and inventories that would measure the important abilities and traits identified via the literature review and expert judgment process. These PTB measures would be piloted and field tested and then revised to become the Trial Battery. See Figure 5 for a flow chart showing the sequencing of the various batteries.) We summarize those findings here. They are more completely reported in Hough, Dunnette, Wing, Houston, and Peterson (1984).

Three types of analyses were done. First, the psychometric characteristics of each scale were explored to pinpoint possible problems with the measures of the construct being measured, so those problems could be avoided when the Pilot Trial Battery measures were developed. These analyses included descriptive statistics, item analyses (including numbers of items attempted in the time allowed), internal consistency reliability estimates, and, for the temperament inventory, percentage of subjects failing the scales intended to detect random or improbable response patterns.

Second, the covariances of the scales within and across the various conceptual domains (i.e., cognitive, temperament, biographical data, and vocational interest) were investigated to detect excessive redundancy among the PB measures, especially across the domains. If such redundancies were detected, then steps could be taken to avoid such a problem in the Pilot Trial Battery. Third, the covariances of the PB scales with ASVAB measures were studied to identify any PB constructs that showed excessive redundancy with ASVAB constructs--again, so that steps could be taken to alleviate such problems for the Pilot Trial Battery. Correlation matrices and factor analyses were the major methods of analysis for these second and third purposes.

The psychometric analyses showed some problems with the cognitive test. The time limits appeared too stringent for several tests, and one test appeared to be much too difficult for the population being tested. Since most of the cognitive tests used in the Preliminary Battery had been developed on college samples or other samples somewhat better educated than the population seeking entry into the Army, these findings were not unexpected.

The lesson learned was that the Pilot Trial Battery measures needed to be accurately targeted (in difficulty of items and time limits) toward the population of persons seeking entry into the Army. No serious problems were unearthed for the temperament, bio-data, and interest scales. Item-total correlations were acceptably high and in accordance with prior findings, and score distributions were not excessively skewed or different from expectation.

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

Correlations and factor analysis of the ten ASVAB subtests and the eight PB cognitive tests confirmed prior analyses of the ASVAB (Kass, et al., 1983) and the relative independence of the PB tests. Although some of the ASVAB-PB test correlations were fairly high (the highest was .57), most were less than .30 (49 of the 80 correlations were .30 or less, 65 were .40 or less). The factor analysis (principal factors extraction, varimax rotation) of the 18 tests showed all eight PB cognitive tests loading highest on a single factor, with none of the ASVAB subtests loading highest on that factor. The non-cognitive scales overlapped very little with the four ASVAB factors identified in the factor analysis of the ASVAB subtests and PB cognitive tests. Median correlations of non-cognitive scales with the ASVAB factors, computed within the five non-cognitive factors, ranged from .03 to .32, but 14 of the 20 median correlations were .10 or less.

Computer Battery Development

Compared to the paper-and-pencil measurement of cognitive abilities and the major non-cognitive variables (temperament, biographical data, and vocational interests), the computerized measurement of psychomotor and perceptual abilities was in a relatively primitive state of knowledge. Much work had been done in World War II using electro-mechanical apparatus, but relatively little work had occurred since then . Microprocessor technology held out the promise of revolutionizing measurement in this area, but the work was (and still is) in its early stages. It was clear, however, that cognitive ability testing was moving into a computer-assisted environment through the methodology of adaptive testing. As Project A began, work was under way to implement the ASVAB via computer-assisted testing methods in the Military Entrance Processing Stations. Therefore, it was also sensible from a practical point of view to investigate these methods of testing.

Roughly speaking, four phases of activities led up to the development of computerized predictor measures for the Pilot Trial Battery: (1) information

gathering about past and current research in perceptual/psychomotor measurement and computerized methods of testing such abilities; (2) construction of a demonstration computer battery, and a continuation of information gathering; (3) selection of commercially available microprocessors and peripheral devices, writing of software for testing several abilities using this hardware, and try out of this hardware and software; and, (4) continued development of software, and design and construction of a custom-made peripheral device, which we called a response pedestal.

We can only mention a few of the high points about this part of the research. Our visits to military laboratories that were then conducting computerized testing taught us that large-scale testing on microprocessors could be accomplished, that a variety of computer languages was in use, that it would be highly desirable to have the computerized test battery be as completely self-administering as possible, and that little information was then available on the reliability or criterion-related validity of computerized measures--because of the recency of their development. By immediately developing a demonstration battery of five tests, we convinced ourselves that some computer languages did not allow enough power and control of timing events for our purposes. We ventured into the area of portable computers, then in its infancy, and found machines that appeared adequate for our needs; namely powerful enough, but also rugged enough to withstand frequent shipping from one field site to another.

We developed our software as "command processors," thus allowing project scientist with no computer language facility to construct entire tests, view and try out items, and revise the tests. Finally, we concluded that responses made through standard key boards and with commercially available joysticks were inadequate for our purposes and designed and had built a customized response pedestal.

DEVELOPMENT OF THE PILOT TRIAL BATTERY

Identification of Measures

In March 1984, a meeting of the predictor team and the scientific advisors was held to decide on the measures to be developed for the Pilot Trial Battery. Information from the literature review, expert judgments, initial analyses of the preliminary battery, and the first three phases of computer battery development was presented and discussed. Predictor team staff made recommendations for inclusions of measures and these were evaluated and revised. Figure 7 shows the results of that deliberation process. The names of the tests developed for the Pilot Trial Battery are shown in the right-hand column of Figure 7. This set of recommendations served as the blueprint for the predictor team's test development efforts for the next several months.

Test Writing and Pilot Tests

Following this meeting, we began writing items for all the instruments. When initial versions of the instruments were complete (or, at least, nearly so), we conducted the first pilot test.

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

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

Figure 7. Predictor categories discussed at IPR March 1984, linked to Pilot Trial Battery test names

Data from the tryout were analyzed, and these results guided revision of the instruments. This process was followed through three iterations, for most of the instruments.

Table 1 describes the pilot tests. Note that we included some marker tests of the constructs for which we were developing new tests.

Field Test

After the third pilot test, we took a little more time to analyze the data, revise the instruments, and prepare for a fairly comprehensive field test of the battery. The objectives of this field test were to provide data sufficient to evaluate the psychometric properties of the battery (including test-retest reliability) and its degree of overlap with the ASVAB. In addition, we collected data to allow analysis of practice effects on the computerized measures and faking/fakability on the temperament/biodata and interest inventories.

A sample size of about 250 was available for the psychometric analyses of the Pilot Trial Battery, about 170 for the analyses of overlap with the ASVAB, about 115 for test-retest analyses, about 75 for practice effects on computers, and about 65-115 for the faking/fakability study (in each experimental cell, total N of about 650). Data were collected primarily at Fort Knox, Kentucky, but data for the faking/fakability study were also collected at the Minneapolis Military Entrance Processing Station and Fort Bragg, N.C.

With a few exceptions, the Pilot Trial Battery was psychometrically sound, and appeared to be measuring abilities that overlapped little with the ASVAB, especially in the temperament/biodata and interest domains. The evaluation of practice effects on computerized test scores showed this to be of little concern. Gain scores after practice were no higher for these tests than for those observed on paper-and-pencil tests of cognitive ability given twice over a short period of time (two weeks or so). The gain scores for computerized test scores ranged from nearly zero to about .4 of a standard deviation, averaging about a quarter of a standard deviation.

We reached the following conclusions from the faking/fakability research.

- Soldiers can distort their responses on the temperament/biodata and interest inventories <u>when instructed to do so</u>.
- Special response validity scales on the temperament/biodata inventory do detect such intentional faking on that inventory and can be used to adjust scores on the substantive scales so as to remove most of the effects of intentional faking.
- Those special response validity scales on the temperament/biodata inventory are not sufficiently effective for detecting and adjusting faked scores on the interest inventory.
- Applicants for the U.S. Army did not appear to be distorting their responses (to appear more favorably qualified). Thus, it appears that intentional distortion may not be a significant problem in

Table. 1.

Summary of Pilot Testing Sessions for Pilot Trial Battery

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

Army applicants in the present volunteer Army. We could not, of course, collect any data that would shed light on this problem in a draft situation.

THE TRIAL BATTERY

Development

At the completion of the field tests, we felt we had shown the Pilot Trial Battery to be ready for use in the concurrent validation research. We used the data from the field test to improve the Pilot Trial Battery measures, but we also had to shorten the length of the battery. It required 6.5 hours to administer the entire battery, and only 4 hours of testing time were available.

Three general principles, consonant with the theoretical and practical orientation that had been used since the inception of the project, guided the revision and reduction decisions:

1. Maximize the heterogeneity of the battery by retaining measures of as many different constructs as possible.

2. Maximize the chances of incremental validity and classification efficiency.

3. Retain measures with adequate reliability.

Using all accumulated information, the final decisions were made in a series of meetings attended by the project staff and by the Scientific Advisory Group. Considerable discussion was generated at these meetings, but the group was able to reach a consensus on the reductions and revisions to be made to the Pilot Trial Battery.

Some tests and scales were dropped, some were shortened, and some redundant items asking about soldier demographics were removed. Table 2 shows the array of measures that made up the Trial Battery. (See Peterson (in press) for a complete description of all research activities leading up through the development of the Trial Battery.)

Trial Battery Scores

As earlier described the Trial Battery was administered to the large, concurrent validity sample. We also collected test-retest data (two week interval) on a subset of about 500 soldiers.

A total of seventy scores was generated from the Trial Battery. Forty-three of these came from the non-cognitive inventories (Assessment of Background and Life Experiences (ABLE), the Army Vocational Interest Career Examination (AVOICE), and the Job Orientation Blank (JOB) - which had been included in the AVOICE for the Pilot Trial Battery but was separately administered for the Trial Battery). Six scores came from the six paper-and-pencil, cognitive tests. Twenty-one scores were generated from the ten

Table 2.

Description of Measures in the Trial Battery

COGNITIVE PAPER-AND-PENCIL TESTS	Number of Items	Time Limit (minutes)
Reasoning Test	30	12
Object Rotation Test	90	7.5
Orientation Test	24	10
Maze Test	24	5.5
Map Test	20	12
Assembling Objects Test	32	16

COMPUTER-ADMINISTERED TESTS	Number of Items	Approximate Time
Demographics	2	4
Reaction Time 1	15	2
Reaction Time 2	30	3
Memory Test	36	7
Target Tracking Test 1	18	8
Perceptual Speed and Accuracy Test	36	6
Target Tracking Test 2	18	7
Number Memory Test	28	10
Cannon Shoot Test	36	7
Target Identification Test	36	4
Target Shoot Test	30	5

NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES	Number of Items	Approximate Time
Assessment of Background and Life Experiences (ABLE)	209	35
Army Vocational Interest Career Examination (AVOICE)	176	20

computer-administered tests. With regard to the computer-administered tests, we did evaluate a number of alternative methods of scoring these tests - such as the use of slopes, intercepts, and slightly different methods of computing means (priority, different methods of trimming items prior to computation of means.) We selected, generally speaking, the most reliable and straightforwardly interpreted scores.

Table 3 shows N's, Means, SD's, reliabilities, and uniqueness (from ASVAB) coefficients for scores on the cognitive, paper-and-pencil tests. Tables 4 and 5 show similar data for the computer-administered tests. Tables 6, 7, and 8 show similar data for the ABLE, AVOICE, and JOB scale scores. (Uniqueness coefficients are not shown for these instruments, but range from .40 to .88, with median U2's of .79 for ABLE, .80 for AVOICE, and .57 for JOB).

As these tables show, the battery possesses adequate to excellent psychometric properties, with the exception of some low reliabilities on a few computer-administered test scores. These low reliabilities primarily occur on the proportion correct scores, and this was anticipated. The items on these tests can almost always be answered correctly if the examinee takes enough time. This operates to severely restrict the range on the proportion correct scores, but increases the variance (and reliability) on the decision time scores, as was our intention.

These Trial Battery scores were the raw material for the validation analyses of the concurrent validity sample, on the "new predictor" side of the equation.

To conclude, we return to the research objectives stated at the beginning of the paper.

1. Identify "best bet" measures -- This objective has been met. As noted, we sifted through a mountain of literature, translating the information onto a common form that enabled us to evaluate constructs and measures in terms of several psychometric and pragmatic criteria. The results of that effort fed into the expert judgment process wherein 35 personnel psychologists provided the data necessary to develop our first model of the predictor space. After further review by experienced researchers in the Army and an advisory group, a set of "best bet" constructs was settled on. We also made some field visits to observe combat arms jobs first-hand, in addition to receiving criterion-side information from other Project A researchers; all of this information was very useful in developing new measures.

2. Develop measures of "best bet" predictors -- This objective was accomplished by following the blueprint provided from the first objective. We carried out many small and not-so-small sample tryouts of these measures as they were developed. The Trial Battery is the tangible product of meeting this objective.

3. Develop procedures for efficiently administering predictor measures -- As anyone who has done research in military settings is aware, soldiers' time is precious and awarded research time is not to be

Table 3.

Concurrent Validity Data Analysis: Means, Standard Deviations, and Reliability and Uniqueness Estimates for Paper-and-Pencil Cognitive Tests

Test	N	Mean	SD	Split-Half Reliability[1]	Test-Retest Reliability[2]	Uniqueness Estimate
Assembling Objects	9343	23.29	6.71	.91	.70	.65
Object Rotation	9345	62.38	19.06	.99	.72	.81
Maze Test	9344	16.39	4.77	.96	.70	.74
Orientation Test	9341	11.02	6.18	.89	.70	.60
Map Test	9343	7.67	5.51	.90	.78	.46
Reasoning Test	9332	19.07	5.67	.87	.65	.53

[1] Split-half reliability estimates were calculated using the odd-even procedure with the Spearman-Brown correction for test length.

[2] Test-Retest reliability estimates are based on a sample of 468 to 487 subjects.

Table 4.

Concurrent Validity Data Analysis: Means, Standard Deviations, and Reliability and Uniqueness Estimates for Computerized Psychomotor Tests

	N	Mean	SD	Odd-Even Reliability	Test-Retest Reliability[1]	Uniqueness Estimate
Target Tracking 1:						
Mean Log (Distance + 1)	9251	2.98	.49	.98	.74	.82
Target Tracking 2:						
Mean Log (Distance + 1)	9239	3.70	.51	.98	.85	.79
Target Shoot:						
Mean Log (Distance + 1)	8892	2.17	.24	.74	.37	.70
Mean Time to Fire	8892	235.39	47.78	.85	.58	.78
Cannon Shoot:						
Mean Absolute Time Discrepancy	9234	43.94	9.57	.65	.52	.56

Note: Time-to-fire and time-discrepancy measures are in hundredths of seconds. Logs are natural logs.

[1] Test-Retest reliability estimates are based on sample sizes of 468 to 487.

Table 5.

Concurrent Validity Sample: Means, Standard Deviations, and Reliability and Uniqueness Estimates for Computerized Cognitive Perceptual Tests

	N	Mean[1]	SD	Split-Half Reliability	Test-Retest Reliability[2]	Uniqueness Estimate
Simple Reaction Time:						
Decision Time Mean	9255	31.84	14.82	.88	.23	.87
Proportion Correct	9255	0.98	0.04	.46	.02	.44
Choice Reaction Time:						
Decision Time Mean	9269	40.93	9.77	.97	.69	.93
Proportion Correct	9269	0.98	0.03	.57	.23	.55
Choice DT Mean - Simple DT Mean[3]	9250	9.09	14.44	.85	.31	.84
Short Term Memory:						
Decision Time Mean	9149	87.72	24.03	.96	.66	.93
Proportion Correct	9149	0.89	0.08	.60	.41	.55
Perceptual Speed & Accuracy:						
Decision Mean Time	9244	236.91	63.38	.94	.63	.92
Proportion Correct	9244	0.87	0.08	.65	.51	.61
Target Identification:						
Decision Time Mean	9105	193.65	63.13	.97	.78	.83
Proportion Correct	9105	0.91	0.07	.62	.40	.59
Number Memory:						
Final Response Time Mean	9099	160.70	42.63	.88	.62	.67
Input Response Time Mean	9099	142.84	55.24	.95	.47	.85
Operations Pooled Mean[3]	9099	233.10	79.72	.93	.73	.66
Proportion Correct	9099	.90	.09	.59	.53	.39
SRT-CRT-STM-PSA-TID:						
Pooled Mean Movement Time[3]	8962	33.61	8.03	.74	.66	.71

[1] Times are given in hundredths of seconds. Logs are natural logs.

[2] N = 460 - 479 for test-retest correlations.

[3] Coefficient Alpha reliability estimates.

Table 6.

ABLE Scale Statistics for Total Group[1]: Trial Battery (Revised)

	No. Items	N	Mean	SD	Median Item-Total Correlation	Internal Consistency Reliability (Alpha)	Test-Retest Reliability[2]
ABLE Substantive Scales:							
Emotional Stability	17	8522	39.0	5.45	.39	.81	.74
Self-Esteem	12	8472	28.4	3.70	.39	.74	.78
Cooperativeness	18	8494	41.9	5.28	.39	.81	.76
Conscientiousness	15	8504	35.1	4.31	.34	.72	.74
Nondelinquency	20	8482	44.2	5.91	.36	.81	.80
Traditional Values	11	8461	26.6	3.72	.36	.69	.74
Work Orientation	19	8498	42.9	6.06	.41	.84	.78
Internal Control	16	8485	38.0	5.11	.39	.78	.69
Energy Level	21	8488	48.4	5.97	.38	.82	.78
Dominance	12	8477	27.0	4.28	.44	.80	.79
Physical Condition	6	8500	14.0	3.04	.60	.84	.85
ABLE Response Validity Scales:							
Unlikely Virtues	11	8511	15.5	3.04	.34	.63	.63
Self-Knowledge	11	8508	25.4	3.33	.36	.65	.64
Non-Random Response	8	8559	7.7	0.59			.30
Poor Impression	23	8492	1.5	1.85	.20	.63	.61

[1] Total group after screening for missing data and random responding.

[2] N = 408 - 412 for test-retest correlation (N = 414 for Non-Random Response test-retest correlation).

[3] Unedited data.

Table 7.

AVOICE Scale Statistics for Total Group[1] (Revised)

AVOICE Scale	No. Items	N	Mean	SD	Median Item-Scale Correlation	Internal Consistency Reliability (Alpha)	Test-Retest Reliability[2]
Clerical/Administrative	14	8463	39.6	10.81	.67	.92	.78
Mechanics	10	8382	32.1	9.42	.80	.94	.82
Heavy Construction	13	8488	39.3	10.54	.68	.92	.84
Electronics	12	8359	38.4	10.22	.70	.94	.81
Combat	10	8466	26.5	8.35	.65	.90	.73
Medical Services	12	8364	36.9	9.54	.68	.92	.78
Rugged Individualism	15	8396	53.3	11.44	.58	.90	.81
Leadership/Guidance	12	8444	40.1	8.63	.62	.89	.72
Law Enforcement	8	8471	24.7	7.37	.65	.89	.84
Food Service - Professional	8	8472	20.2	6.50	.67	.89	.75
Firearms Enthusiast	7	8397	23.0	6.36	.66	.89	.80
Science/Chemical	6	8468	16.9	5.33	.70	.85	.74
Drafting	6	8493	19.4	4.97	.66	.84	.74
Audiographics	5	8473	17.6	4.09	.69	.83	.75
Aesthetics	5	8413	14.2	4.13	.59	.79	.73
Computers	4	8224	14.0	3.99	.78	.90	.77
Food Service - Employee	3	8304	5.1	2.08	.54	.73	.56
Mathematics	3	8421	9.6	3.09	.78	.88	.75
Electronic Communication	6	8403	18.4	4.66	.60	.83	.68
Warehousing/Shipping	2	8407	5.8	1.75	.44	.61	.54
Fire Protection	2	8431	6.1	1.96	.62	.76	.67
Vehicle/Equipment Operator	3	8378	8.8	2.65	.51	.70	.68

[1] Total group after screening for missing data and random responding.

[2] N = 389 - 409 for test-retest correlation.

Table 8.

JOB Scale Statistics for Total Group[1] (Revised)

(Trial Battery)

JOB Scale	No. Items	N	Mean	SD	Median Item-Total Correlation	Internal Consistency Reliability (Alpha)
Organizational Support	10	7809	43.6	4.51	.54	.84
Job Status	5	7817	21.6	2.33	.43	.67
Serve Others	3	7784	12.1	1.83	.52	.66
Autonomy	4	7817	15.1	2.29	.31	.50
Routine	4	7707	9.6	2.30	.25	.46
Ambition	3	7751	12.4	1.63	.35	.49

[1] Total group after screening for missing data and random responding.

squandered. We think we have developed and implemented effective methods for getting maximum quality and quantity of data out of our data collection efforts. The favorable results we have so far achieved in completeness and usefulness of data are due in large part, we think, to the attention paid to this objective.

4. Estimate reliability and vulnerability of measures -- This objective has also been largely accomplished. Analyses to date indicate that the new measures are psychometrically sound and acceptably invulnerable to the various sources of measurement problems -- or we have devised some ways to adjust for such effects. However, more specifically targeted research would be useful in this area.

5. Determine the interrelationships between the new measures and current preinduction measures -- Work still remains on this objective, but the data collected to date show that the new measures have much variance that is not shared with the ASVAB, and that the across-domain shared variance is low (e.g., the new cognitive measures have low correlations with the non-cognitive measures).

6. Determine the level of prediction of solider performance, classification efficiency, and incremental validity of the new and measures -- alas, other presenters at this symposium are providing this information, so I will now shut up and sit down.

References

Alley, W. E., & Matthews, M. D. (1982). The vocational interest career examination. Journal of Psychology, 112, 169-193.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Flanagan, J. C. (1965). Flanagan industrial test manual. Chicago: Science Research Associates.

Hough, L. M., Dunnette, M. D., Wing, H., Houston, J. S., & Peterson, N. G. (1984). Covariance analyses of cognitive and non-cognitive measures of Army recruits: An initial sample of Preliminary Battery Data. Presented at the 92nd Annual Convention of the American Psychological Association, Toronto. In Eaton et al. (eds.) (1984). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1984 fiscal year (ARI Technical Report 660). Alexandria, VA: Army Research Institute.

Hough, L. M., Kamp, J. D., & Barge, B. N. (1988). Utility of temperament, biodata, and interest assessment for predicting job performance: a review and integration of the literature. ARI Research Note, ADA 178944

Jackson, D. N. (1967). Personality Research Form Manual. Goshen, NY: Research Psychologists Press.

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1983). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10: 1981 Army applicant sample. Educational and Psychological Measurement, 43, 1077-1088.

McHenry J. J., & Rose, S. R. (1988). The validity and potential usefulness of psychomotor ability tests for personnel selection and classification. ARI Research Note, 88-13 ADA 193558.

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. Journal of Applied Psychology Monographs, 64, 569-607.

Peterson, N. G. (1987). Development and field test of the Trial Battery for Project A. ARI Technical Report.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80, (1, Whole No. 609).

Ruch, F. L., & Ruch, W. W. (1980). Employee Aptitude Survey: Technical Report. Los Angeles, CA: Psychological Services, Inc.

Schmidt, F. L., Hunter, J. E., Croll, P. R., & KcKenzie, R. C. (1983). Estimation of employment test validities by expert judgment. Journal of Applied Psychology, 68, 590-601.

Toquam, J. L., Corpe, V. A., & Dunnette, M.D. (In press). Cognitive abilities: A review of theory, history, and validity. ARI Research Note.

Wing, H., Peterson, N.G., & Hoffman, R. E. (1984). Expert judgments of predictor-criterion validity relationships. Presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Ontario, Canada. In Eaton et al. (eds.) (1984). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year (ARI Technical Report 660). Alexandria, VA: Army Research Institute.

AN EXAMINATION OF RACE AND SEX EFFECTS ON PERFORMANCE RATINGS

Elaine D. Pulakos
American Institutes for Research

Leonard A. White
U.S. Army Research Institute

Walter C. Borman
Personnel Decisions Research Institute

Presented at the Annual Conference of the
Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Abstract

This research investigated the effects of rater source (peer, supervisor), rater and ratee race (black, white, Hispanic), rater and ratee sex, and job type on ratings collected for 39,537 Army enlisted personnel. The results showed that race and sex did not interact in their effects on the ratings. Although significant effects were observed for sex, race, rater source, and job type, the proportions of variance accounted for by these effects were minimal. Of particular interest was that the race effects found here were considerably less than those reported in a recent meta-analysis (Kraiger & Ford, 1985). Results and implications are discussed.

An Examination of Race and Sex Effects on Performance Ratings

Considerable research has investigated rater and ratee gender and/or race effects on ratings. Unfortunately, inconsistent findings have resulted from this body of literature. Regarding gender effects, the only consistent research finding is that rater and ratee sex do not appear to interact in their effects on evaluative judgments (e.g., Bartol & Butterfield, 1976; Mobley, 1982; Pulakos & Wexley, 1983). Although significant main effects for sex are evident, many of the studies reporting such effects have been conducted in a laboratory. Because relatively little field research has investigated gender effects on ratings, it is difficult to draw definitive conclusions about the existence or lack thereof of sex effects in ongoing performance appraisal situations.

Inconsistent findings have also resulted regarding rater and ratee race effects on performance ratings. However, a recent meta-analysis of ratee race effects revealed corrected mean correlations between ratee race and ratings given by black and white raters of .183 and -.220, respectively, indicating that both black and white raters assigned significantly higher ratings to ratees of their own race than to ratees of the other race (Kraiger & Ford, 1986). The meta-analysis also showed that ratee race effects were more likely to be found in field studies in which blacks constituted a small percentage of the workforce.

The present research investigated the effects of rater and ratee gender/race on ratings. With few exceptions, previous field research has lacked adequate sample sizes to support an investigation of race x sex interactions. In addition, this research extends previous investigations of race and sex effects on ratings in two important ways.

First, using a common set of rating scales, both peers and supervisors rated thousands of ratees occupying 19 different jobs. It was thus possible to examine potential differences in race and sex effects as a function of the rating source (peer or supervisor) and the type of job held by the ratee. Second, three levels of rater and ratee race (i.e., whites, blacks, and Hispanics) were included rather than including only whites and blacks, as has been characteristic of the vast majority of research investigating race effects.

Method

Sample

The data reported here were collected as part of Project A, the Army's multi-year research program to develop an improved selection and classification system for enlisted personnel. A total of 6377 supervisors and 8174 peers evaluated first-term soldiers representing 19 jobs selected to be representative of the entire population of Army jobs. The supervisors rated an average of 2.29 subordinates, and the peers rated an average of 3.05 co-workers, yielding a total sample of 39,537 rater-ratee pairs. Table 1 shows a breakdown of the number of pairs representing each rater/ratee race and rater/ratee sex combination.

Procedure

Peer and supervisor ratings were collected on 10 7-point behavioral rating dimensions that were developed to assess first-term soldier effectiveness in any Army job. Raters were trained on how to use the rating scales properly and on how to avoid several common rating errors.

A principal components analysis with a varimax rotation was used to identify constructs underlying the performance ratings. A three-factor solution was chosen as the most psychologically meaningful, with the factors named and defined as shown in Table 2. For each ratee, three unweighted composites were calculated using the dimension ratings that had the highest loadings on each factor. Alpha coefficients for the composite scores were: Technical Skill and Job Effort (.88), Personal Discipline (.80), and Military Bearing (.62). Intercorrelations among the composites ranged from .51 to .74.

Results

A preliminary multivariate analysis of variance (MANOVA) revealed that race and sex did not interact in their effects on the three composites. Hence, race and sex effects were examined separately in subsequent analyses. Preliminary MANOVAs were also performed to examine job-type x race and job-type x sex interactions. Although significant interactions resulted, the rating variance accounted for by these was minimal (i.e., less than one-half of one percent). Thus, job-type was excluded from further analyses.

Race Effects on the Ratings

A 2 (Rating Source) x 3 (Rater Race) x 3 (Ratee Race) MANOVA was conducted to examine race effects on the three rating measures. The levels of rating source were peer and supervisor, and the levels of race were black, white, and Hispanic. Upon obtaining a significant MANOVA, univariate analyses were examined for each dependent measure. These results are shown in Table 3. The means for all rater group by race combinations are shown in Table 4.

With the exception of the ratee race main effect for the Bearing factor, individual effects accounted for substantially less than one percent of the rating variance. In fact, the total proportions of variance accounted for by all rater source and race effects were quite small (i.e., Technical Skill and Job Effort, r^2 = .016; Personal Discipline, r^2 = .003; and Military Bearing, r^2 = .028). Because of the minimal variance accounted for, interpretation of the interactions will not be undertaken. It is interesting to note, however, that the nature of the effects was not consistent across the three rating factors. For instance, blacks were rated higher than whites on Military Bearing but lower than whites on the other two dimensions.

Sex Effects on the Ratings

A 2 (Rating Source) x 2 (Rater Sex) x 2 (Ratee Sex) MANOVA was conducted to examine sex effects on the ratings. Again, upon obtaining a significant MANOVA, univariate analyses were examined. These results are presented in Table 5. The means for each rating source by sex combination are shown in Table 6.

Similar to the race analyses, the proportions of rating variance accounted for by the significant effects were minimal. The total proportions of variance accounted for by all rater source and sex effects were as follows: Technical Skill and Job Effort (r^2 = .012), Personal Discipline (r^2 = .001), and Military Bearing (r^2 = .004). In addition, the directions of the significant main effects and interactions were not consistent across the three rating factors.

Repeated Measures Analyses

To determine how the results reported above may have been affected by the fact that the rating observations were not independent (i.e., raters rated multiple ratees), a 2 x 2 x 2 repeated measures MANOVA was conducted with rater source (peer or supervisor) and rater race (black or white) constituting the between subjects factors, ratee race (black or white) as the single within subjects factor, and measures of the three rating factors as the multiple dependent measures. Unfortunately, sufficient data were not available to include Hispanics in this analysis. A repeated measures MANOVA like that described above was also conducted to investigate the sex effects. These analyses yielded results virtually identical to those reported above. The only exception was that in the repeated measures analysis, the two- and three-way interactions involving ratee sex were nonsignificant for all three dependent measures.

Discussion

The present field research investigated race and sex effects on peer and supervisor ratings of ratees occupying a variety of jobs. The overwhelming finding was that the proportions of variance accounted for by gender and, especially, race were less than have been found in previous research. For example, Kraiger and Ford (1985) reported correlations between ratee race and ratings for black and white raters of .183 and -.220, with the variance accounted for by these correlations equal to three and five percent, respectively. The present variance accounted for by, especially, the race by ratee race interactions was substantially less than one percent.

One difference between the Kraiger and Ford (1985) research and this research is that no corrections (e.g., for unreliability) were made here. In order to enable a more direct comparison between Kraiger and Ford's results and those reported here, a meta-analysis similar to Kraiger and Ford's was conducted. The proportion of rating variance accounted for by ratee race was still much less than reported in Kraiger and Ford's research. Thus, while we believe that future research should focus on possible explanations for observed effects rather than on effect sizes alone, it may be premature to accept Kraiger and Ford's analysis results as the best estimate of the ratee race effect size in the population.

One explanation for the present race results is that raters were trained to focus specifically on ratee job performance and to avoid using nonperformance factors (e.g., sex, race) as a basis for their evaluations. Another possible explanation is that racial bias may be less prevalent in military versus civilian work settings due to reasonably large percentages of minority service members. This explanation is consistent with Kraiger and Ford's (1985) finding that race effects were less likely to be found when blacks constituted a larger percentage of the workforce.

Because no meta-analysis has been conducted to estimate population sex effect sizes on ratings, it is more difficult to compare the magnitudes of the present sex effects to previous research. Further, relatively few field studies have reported the rating variance accounted for by gender. Nevertheless, in some cases, reported effect sizes have

been larger than those found here (e.g., Mobley, 1982), whereas in other cases, sex has been shown to have no effect on ratings (e.g., Thompson & Thompson, 1985) or to account for only minimal proportions of the rating variance (Pulakos & Wexley, 1983).

Two additional points are worth mentioning regarding the results of this research. First, Landy and Farr (1980) concluded that sex stereotype of the occupation appears to interact with ratee sex such that males receive more favorable evaluations than females in traditionally masculine occupations but that no differences or smaller differences in favor of females occur in traditionally feminine occupations. Although significant job type x ratee sex interactions were observed in this study, the proportions of the variance accounted for by these effects were trivial. *Beyond* this, however, even the nature of the significant effects did not provide support for Landy and Farr's sex-role stereotype hypothesis.

The second noteworthy point concerns the lack of sex x race interactions found in this study. Because of inadequate sample sizes, most performance appraisal field research has been unable to investigate whether or not race and sex interact in their effects on ratings (see Thompson & Thompson, 1985 for an exception). There has, however, been some assessment center research (e.g., Huck & Bray, 1976; Schmitt & Hill, 1977) in which significant race x sex interactions have been observed. As an example, in the Schmitt and Hill study, black females tended to be rated lower when they were in assessment groups with larger proportions of white males. The results of the present study along with

nonsignificant race x sex interactions reported by Thompson and Thompson (1985) seem to suggest that the interactive effects of race and sex found in assessment center ratings do not generalize to performance appraisal situations. It may be that because assessment centers are characterized by relatively short durations of interpersonal contact between assessors and assessees as well as a more limited amount of ratee performance information, cues of race and sex may be more salient to assessors, increasing the probability that these factors will have greater influence on the ratings (Wendelken & Inn, 1981).

Although the present research clearly shows that systematic bias as a function of rater or ratee sex and race was not an important factor influencing the ratings, future research could examine the evaluation processes involved when the same versus different race or sex raters evaluate ratees. For example, irrespective of whether or not there are mean subgroup differences in ratings, raters may use different cues when evaluating someone of a different race or sex versus a person of the same race or sex. Policy capturing (e.g., Zedeck & Kafry, 1977) or a lens model approach (Schmitt, Noe, & Gottschalk, 1986) are possible strategies for investigating such similarities and differences.

References

Bartol, K. M., & Butterfield, D. A. (1976). Sex effects in evaluating leaders. Journal of Applied Psychology, 61, 446-454.

Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of white and black females. Personnel Psychology, 29, 13-30.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. Journal of Applied Psychology, 70, 56-65.

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

Mobley, W. H. (1982). Supervisor and employee race and sex effects on performance appraisals: A field test of adverse impact and generalizability. Academy of Management Journal, 25, 598-606.

Pulakos, E. D., & Wexley, K. N. (1983). The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. Academy of Management Journal, 26, 129-139.

Schmitt, N., & Hill, T. (1977). Sex and race composition of assessment center groups as a determinant of peer and assessor ratings. Journal of Applied Psychology, 62, 261-264.

Schmitt, N., Noe, R. A., & Gottschalk, R. (1986). Using the lens model to magnify raters' consistency, matching, and shared bias. Academy of Management Journal, 29, 130-139.

Thompson, D. E., & Thompson, T. A. (1985). Task-based performance appraisal for blue-collar jobs: Evaluation of race and sex effects. Journal of Applied Psychology, 70, 747-753.

Wendelken, D. J., & Inn, A. (1981). Nonperformance influences on performance evaluations: A laboratory phenomenon? Journal of Applied Psychology, 66, 149-158.

Zedeck, S., & Kafry, D. (1977). Capturing rater policies for processing evaluation data. Organizational Behavior and Human Performance, 18, 269-294.

Table 1

Breakdown of the Rater/Ratee Pairs by Race and Sex Composition

Rater Race/Ratee Race	N	Rater Sex/Ratee Sex	N
Black/Black	4,700	Male/Female	2,437
Black/White	7,391	Male/Male	20,773
Black/Hispanic	502	Female/Female	1,273
Hispanic/Black	617	Female/Male	1,686
Hispanic/White	1,365		
Hispanic/Hispanic	115		
White/Black	5,745		
White/White	18,294		
White/Hispanic	808		
Total	39,537		26,169

Note. The total number of dyads for the different sex combinations is smaller than the total number of dyads for the race combinations because five of the 19 MOS were combat jobs in which there were no females. While the percentages of the total sample represented by different race and sex combinations are variable, they nevertheless accurately represent the corresponding percentages found in the Army population.

Table 2

Army-Wide Factors

Name	Definition
Technical Skill and Job Effort	Exerting effort and showing proficiency over the full range of job tasks; engaging in training or other developmental activities to increase proficiency; persevering under dangerous or adverse conditions; and demonstrating leadership and support towards peers.
Personal Discipline	Adhering to Army rules and regulations; exercising self-control; demonstrating integrity in day-to-day behavior; and, not causing disciplinary problems.
Military Bearing	Maintaining an appropriate military appearance and bearing and staying in good physical condition.

Table 3

Analysis of Variance Results for Race

Effect	df	Technical Skill and Job Effort F	Technical Skill and Job Effort Nature of Main Effects	Personal Discipline F	Personal Discipline Nature of Main Effects	Military Bearing F	Military Bearing Nature of Main Effects
Rating Source (A)	1	92.61*	P>S	19.54*	P>S	7.75*	S>P
Rater Race (B)	2	31.26*	B,H>W	8.45*	B,H>W	3.32*	B>W
Ratee Race (C)	2	47.27*	H>W>B	15.24*	H,W>B	209.94*	B,H>W
A x B	2	5.17*		8.49*		2.59*	
A x C	2	1.69		1.99		1.25	
B x C	4	28.99*		3.87*		6.69*	
A x B x C	4	3.10*		1.80		2.67*	

Note. *$p<.05$. Regarding the Rating Source main effects: P = peer raters and S = supervisor raters. Regarding the Rater and Ratee Race main effects: B = black, W = white, H = Hispanic.

Table 4

Means and Standard Deviations for Rater Group By Race Combinations

	Technical Skill and Job Effort		Personal Discipline		Military Bearing	
	Peer	Supervisor	Peer	Supervisor	Peer	Supervisor
White Rater						
White Ratee	4.47	4.29	4.53	4.54	4.70	4.77
	(1.09)	(1.18)	(1.21)	(1.28)	(1.20)	(1.24)
Black Ratee	4.20	3.95	4.40	4.42	5.07	5.15
	(1.14)	(1.20)	(1.26)	(1.32)	(1.13)	(1.17)
Hispanic Ratee	4.39	4.24	4.58	4.54	4.93	5.06
	(1.03)	(1.12)	(1.17)	(1.25)	(1.17)	(1.17)
Black Rater						
White Ratee	4.44	4.33	4.56	4.56	4.60	4.79
	(1.07)	(1.17)	(1.22)	(1.32)	(1.24)	(1.30)
Black Ratee	4.53	4.19	4.59	4.49	5.16	5.25
	(1.06)	(1.16)	(1.20)	(1.31)	(1.14)	(1.17)
Hispanic Ratee	4.66	4.44	4.76	4.68	4.98	5.26
	(0.94)	(1.11)	(1.21)	(1.27)	(1.13)	(1.25)
Hispanic Rater						
White Ratee	4.63	4.18	4.72	4.44	4.76	4.62
	(1.04)	(1.23)	(1.19)	(1.36)	(1.19)	(1.33)
Black Ratee	4.39	4.02	4.52	4.34	5.03	5.16
	(1.07)	(1.22)	(1.26)	(1.44)	(1.17)	(1.21)
Hispanic Ratee	4.99	4.41	5.13	4.42	5.28	5.27
	(0.83)	(1.03)	(1.01)	(1.33)	(0.89)	(1.09)

Table 5

Analysis of Variance Results for Sex

Effect	df	Technical Skill and Job Effort F	Nature of Main Effects	Personal Discipline F	Nature of Main Effects	Military Bearing F	Nature of Main Effects
Rating Source (A)	1	53.68*	P>S	2.24		25.21*	S>P
Rater Sex (B)	1	4.30*	F>M	3.54		20.74*	F>M
Ratee Sex (C)	1	14.70*	M>F	.08		23.20*	M>F
A x B	1	2.04		2.20		.43	
A x C	1	.29		.12		.96	
B x C	1	3.14		16.43*		8.79*	
A x B x C	1	10.03*		3.79		.36	

Note. *p<.05. Regarding the Rating Source main effects: P = peer raters and S = supervisor raters. Regarding the Rater and Ratee Sex main effects: M = male and F = female.

Table 6

Means and Standard Deviations for Rater Group By Sex Combinations

	Technical Skill and Job Effort		Personal Discipline		Military Bearing	
	Peer	Supervisor	Peer	Supervisor	Peer	Supervisor
Female Rater						
Male Ratee	4.54	4.36	4.62	4.60	5.00	5.14
	(1.05)	(1.16)	(1.19)	(1.32)	(1.15)	(1.20)
Female Ratee	4.46	4.14	4.56	4.40	4.76	4.94
	(1.02)	(1.12)	(1.22)	(1.26)	(1.20)	(1.28)
Male Rater						
Male Ratee	4.47	4.22	4.56	4.53	4.81	4.90
	(1.08)	(1.19)	(1.21)	(1.31)	(1.21)	(1.26)
Female Ratee	4.32	4.25	4.65	4.68	4.72	4.88
	(1.10)	(1.18)	(1.19)	(1.35)	(1.21)	(1.29)

DESIGNING, PLANNING, AND SELLING PROJECT A

Joyce L. Shields
Lawrence M. Hanser

U.S. Army Research Institute

Presented at the Annual Conference of the
Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

Like many events, Project A was a product of the people and time of its conception. In this paper we first describe the Zeitgeist which existed prior to and during the planning of this project. Second, we discuss the design of the project as a prime example of successful policy research. Finally, we address the issue of selling the project initially, and maintaining support for long-term research in the face of changing problems and goals.

The Zeitgeist

The events which shaped the Army and eventually resulted in Project A began several years earlier, in the 1970's. More than 14.9 million American youth were drafted between 1940 and 1973 (Nelson, 1983). At the close of the Vietnam War in 1973 the draft came to an end and the All-Volunteer Force (AVF) was born. By 1975 first term attrition had reached 26.6% among high school graduate enlistees and 51.4% among non-high school graduate enlistees, both record highs. Also in that year, only 58% of Army enlistees had a high school diploma, compared with 90% this year. Although the size of the Army had been reduced drastically from the Vietnam War era, these high attrition rates placed an enormous burden on recruiting. These times were best summarized in General Meyer's now famous White Paper (1980) on the 'Hollow Army.'

In addition to changes in the personnel system of the Army, the Army was beginning the largest force modernization program since World War II. Anti-tank weapons were now becoming wire-guided missiles; tanks would have on-board computer systems for gunnery and navigation; infantry squads would use satellite communications for determining their battlefield location; and shoulder fired missiles would include state-of-the-art electronics for aircraft identification. Further complicating the increasing technical demands of modern equipment was the prediction of a significant decline in the number of eligible youth which was projected to begin about 1982 and continue through 1996. Obviously the personnel needs of the Army were facing substantial change in a climate of declining manpower supply.

The climate was also unfavorable to testing. The nation as a whole was questioning the fairness of tests. In 1978 the Uniform Guidelines were published. The Congress, in 1981, issued a directive that the Services must "develop a better database on the relationship between factors such as high school graduation, entrance test scores, age, etc., and potential for effective service." Interest in, and support for testing research in the Army had declined substantially. The Army Research Institute, the traditional home for selection and classification research in the Army, was organized into two laboratories at that time, the Training Research Laboratory (TRL) and the Organization and Systems Research Laboratory (OSRL). OSRL included only a small team of people devoted to selection and classification research.

In 1980, the Armed Services Vocational Aptitude Battery Forms 6/7 (ASVAB 6/7) which was used operationally from 1976 to 1980 was discovered to have been misnormed. As a result of the misnorming, in 1980, 50% of Nonprior Service Army Recruits were drawn from the bottom 30% of the eligible youth population. Today, over 60% of recruits come from the top 50% of the youth population. With this large influx of low-scoring recruits in the late 70's the Army began to question what difference entry test scores made in terms of eventual performance in military occupations. That is, did it really matter whether the Army recruited individuals from a higher percentile in the youth population? Unfortunately, this question could not be adequately addressed, because at the inception of the AVF, the Training and Doctrine Command (TRADOC) introduced criterion-referenced training, go/no go testing, and mastery learning, so that no reasonable criteria existed. Further, Skill Qualification Test scores (i.e., mid-career tests of job knowledge) were not readily accessible, and often, centralized recordkeeping systems, where such information was stored, were cross-sectional rather than longitudinal in nature.

As is now clear, the Army was facing several problems:

Was it possible to demonstrate a relationship between selection tests and performance in military occupations?

Could selection tests be used to identify individuals more likely to complete their tour of service?

Given the declining manpower pool, could tests be designed to more efficiently use the available resources?

Could individuals be better allocated to the diverse demands of the Army and Army occupations?

Could weapon systems be designed, with enhanced battlefield effectiveness, which would match the available skills of the declining pool of operators and maintainers?

These problems cut across a number of Army commands and organizations, so that resolving them was important to a wide variety of senior Army leaders. The project did not spring from a desire to examine the issues related to validity generalization, or rater accuracy, or computerized testing, or a basic desire to support industrial/organizational research. Rather it grew from the need to address some very real policy issues.

The People

According to an Army Science Board report by Alexander (1980), "It is not enough for a research community to exist, or even for it to be working on problems of concern to the policymakers. Strong and intimate links are essential to transmit problems and questions, to convert them into researchable projects, and to transmit the results back to the client -- not as research reports -- but as options, alternatives, and evaluations that the policymaker can use... a special type of researcher is required -- one who understands both the research and the policy worlds... (there is also) a requirement for people on the Army side who are sensitive to the analytical approach and to the potential contributions of research to policymaking."

In 1980, this situation existed. Although a number of such people were in key positions at that time, we would be remiss were we not to mention the presence and support of General Maxwell R. Thurman. In his roles as Commander of the U.S. Army Recruiting Command, Deputy Chief of Staff for Personnel, Vice Chief of Staff of the Army, and now Commander of the Training and Doctrine Command, General Thurman continued to be actively involved in this research.

The Plan

Upon examining the list of problems facing the Army in the late 1970's, it is clear that a number of discrete policy research projects could have been designed to address them. In fact, the tendency is strong for that to happen. However, rather than simply forging ad hoc solutions to the laundry list of problems, a comprehensive program of personnel selection research was established. But it was designed in such a way as to provide the necessary basic data on which to both resolve ad hoc problems as well as to address longer term scientific issues.

In 1981, ARI initiated a multiyear, multimillion dollar research program, consisting of two interrelated projects, to relate better selection and classification measures and procedures to the criterion of soldier performance. The objectives of this program were to:

Validate ASVAB against soldier performance

Develop new selection and classification measures and procedures to optimize the soldier requirements match

Design computer-based decision aides for managers of the Army's manpower processes

At the same time, ARI organized a Manpower and Personnel Research Laboratory (MPRL), which included the Personnel Utilization Technical Area, to be responsible for this program of research. In the Spring of 1981, two teams of individuals from this technical area began to prepare the statements of work which were

to become Project A (development and validation of enlistment measures) and Project B (development of a computer-based system to link personnel requirements with resources), addressing the major objectives outlined above. After several months of writing and rewriting, the Requests for Proposals (RFP) were released in the fall of 1981. A contract for Project A was signed with the Human Resources Research Organization, American Institutes for Research, and Personnel Decisions Research Institute in September 1982.

Project A as Policy Research

According to Alexander (1980), successful policy research has the following characteristics:

Importance: The research should be concerned with important issues.

Crosscutting: Topics chosen for analysis should crosscut organizational boundaries.

Understanding the environment: Researchers need to understand the decision environment and bureaucratic context of the policymaker.

Confidence and trust: Policymakers must have confidence in the technical ability of the researchers and the researchers should view their clients as people who value their efforts.

Accountability: Researchers must be accountable for their products and their results. The research should be available to others for inspection, review, and debate.

Tolerance for wrong answers: The probability of "wrong", ambiguous, and complicated results must be recognized and accepted by clients.

The designing and planning of Project A is related to the chararcteristics described above in the following ways:

Importance. As discussed previously the problems addressed in this research program are of great importance to the Army.

Crosscutting. These problems are of interest to many constituencies, including personnel and training proponents. Both the military and civilian sides of the Army are equally interested in the results, although for different reasons.

Understanding the Environment. The researchers understand the Army and are given/allowed access to top policy makers in the Army. There are many different points of view, and researchers are given the opportunity to question strongly held positions. The researchers are problem oriented and responsive, willing to

work quickly when possible to provide short term answers in exchange for a long term committment on the part of policymakers. Researchers continue to provide information back to policymakers in terms of options, alternatives, and evaluations - not just research reports.

Confidence and trust. Key policymakers have confidence and trust in the technical ability of the researchers and their understanding of the problems. Key policymakers invested, and continue to invest in the researchers, and provide time, access to sensitive data, and all necessary support as well as trust and confidence.

Accountability. The research plan was well founded on a sound scientific base. It was and continues to be open for inspection by both the scientific and policy communities.

Tolerance for wrong answers. The Army clients are not only open to results - whether or not prior beliefs are confirmed, but they have been and continue to be willing to use the results to change and set policy.

The Changing Environment

As researchers, we have a tendency to judge the success of a project by how well we have solved the problems which originally generated it. The list of personnel selection and classification problems which the Army faces today would be somewhat different from the list we mentioned earlier. Policymakers are not interested in solutions to problems which no longer exist, but rather in the problems which they face today. The challenge for Project A, and all such long-term projects, is to continually readjust as policy problems change, so that the research remains relevant to policymakers.

References

Alexander, Arthur (1980). Policy Research on Human Issues. Washington, D.C.: Army Science Board, Human Issues Group #2.

Nelson, Gary R. (1983). The Supply and Quality of First-Term Enlistees Under the All-Volunteer Force. In W. Bowman, R. Little, and G. T. Sicilia (Eds.), The All-Volunteer Force After a Decade. Washington, D.C.: Pergammon-Brassey's International Defense Publishers.

Meyer, Edward C. (1980). The Hollow Army (White Paper). Washington, D.C.: Department of the Army.

PREDICTIVE VALIDITY OF NONCOGNITIVE MEASURES FOR ARMY CLASSIFICATION AND ATTRITION

Hilda Wing

U.S. Army Research Institute

Leaetta M. Hough
Norman G. Peterson

Personnel Decisions Research Institute

Presented at the *Annual Conference of the* Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

Predictive Validity of Noncognitive Measures for Army Classification and Attrition

Abstract

Over 9,000 soldiers in four military occupational specialties were administered vocational interest measures, biographical questionnaires, and temperament surveys as they entered their service careers. Approximately nine months later, follow up research determined whether these soldiers were still in their initial occupations or even still in the Army. Selected vocational interest measures predicted occupational classification fairly well for two of the four occupations; selected biodata and temperament measures predicted early attrition fairly well given the low base rate of this dependent variable.

Predictive Validity of Noncognitive Measures for Army Classification and Attrition

Many (for example, Campbell, 1986) have argued that performance is inherently multidimensional. One way of conceptualizing performance space is to divide it into "can-do" and "will-do" subspaces. The former might consist of the technical skills and abilities indexing the maximum quality and quantity of productivity of which an individual is capable. The latter would then be composed of those attitudes and characteristics indexing the typical performance level of the individual. Cognitive abilities predict the former performance subspaces fairly well; noncognitive measures such as vocational interests, biodata, and temperament measures provide some promise for predicting the latter. It is these typical performance measures which are of concern in this report.

As part of the Army's Project A, a Preliminary Battery of paper and pencil measures was administered to soldiers in four selected Military Occupational Specialties (MOS) as they entered military service during late 1983 and early 1984. The battery included measures of vocational interests, individual history background or biodata, and temperament. These measures were used to predict whether a soldier would still be in the Army some time after initial training, in this case, December, 1984. The average length of time a soldier had been in the service was nine months, with a range from six to eighteen months.

The hypotheses of interest concerned classification and prediction. Hypothesis One concerned the efficacy of vocational interest measures in predicting MOS membership: Members of the four very different MOS should show different average scores on the vocational interest measures. Such a finding would provide support for the "gravitational hypothesis" (McCormick, Jeanneret, & Mecham, 1972) which suggests that people of different interests "gravitate" towards those occupations which they find most compatible. Hypothesis Two concerned the efficacy of all the noncognitive measures in predicting early attrition, whether the tested soldier was still in the service by December, 1984, when the records were evaluated. This hypothesis has two components, one for the biodata and temperament measures and one for the vocational interests. For the former, prior research (Hough, Dunnette, Wing, Houston, & Peterson, 1984) has shown biodata and temperament measures to cover the same constructs or variables. In this instance, the concern was with the attitudinal and socialization variables which might predict whether an individual would be ill-behaved, hence a potential discipline problem, or would have academic difficulty with Army training. For the latter, corroborating evidence would consist of attritees having less compatible interests for a given MOS than the stayers. It is the case that cognitive variables are effective in predicting training criteria; the questions here concerned whether the noncognitive variables could provide predictability in addition to the cognitive variables currently used in Army selection and classification.

Method

Research Participants

The population from which these examinees were selected consisted of those soldiers (recruits) who had entered active duty in the Regular Army and who had begun training in one of four MOS at one of five selected Army posts between October 1, 1983, and June 30, 1984, as follows:

MOS 19A: Tank Crewman. The sample consisted of 2,614 male soldiers.

MOS 31C: Radio Teletype Operator. The sample consisted of 1,989 soldiers, which included 280 females.

MOS 63B: Vehicle and Generator Mechanic. The sample included 2,197 soldiers of whom 129 were female.

MOS 71L: Administrative Specialist. The sample included 2,798 soldiers of whom 1,350 were female.

The groups were ethnically diverse, each having about five percent Hispanics and over twenty percent Blacks, except for the 71L's, of which over forty percent were Black. No analyses were performed separately by race/ethnicity or by sex.

Variables

Predictors

ASVAB. Before entry into military service, each soldier had taken the Armed Services Vocational Aptitude Battery (ASVAB), a 3 1/2 hour cognitive test battery used for selection and classification into all the military services All recruits had to achieve a minimum score on a composite known as the Armed Forces Qualification Test (AFQT), summed from scores on four subtests. High school graduates had to be at or above the 21st percentile, while nongraduates had to be at or above the 31st percentile, based on World War II norms for male military personnel. Second, each MOS had a specific composite on which a minimal score was required for entry. These minima were roughly equivalent (McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984) to the 26th percentile of the AFQT for Armor Crewman and Mechanic, and to the 39th percentile for Radio Teletype Operator and Administrative Specialist.

Preliminary Battery (PB). The PB required about four hours to administer. It included eight spatial/perceptual measures which will not be discussed further here. Also included were the 18 scales from the Air Force Vocational Interest Career Examination (VOICE; Alley & Matthews, 1982); five temperament scales adapted from published scales [two from the Differential Personality Questionnaire or DPQ (Tellegen, 1982); one from the California

Psychological Inventory or CPI (Gough, 1975); the Rotter I/E Scale (Rotter, 1966), and validity scales from both the DPQ and the Personality Research Form or PRF (Jackson, 1967)]; and Owens' (Owens & Schoenfeldt, 1979) Biographical Questionnaire (BQ). The BQ was scored for 22 scales based on prior analyses of an initial sample (Hough et al., 1984). Items tapping religion or socioeconomic status had been deleted while items tapping curricula, coursework, and physical fitness had been added. These prior analyses had determined the structure of these sex-independent scales.

The names of the scales used in the analysis reported here, with their numbers of items, are as follows. The range and median values of coefficient alpha for each set are also given. Each measure was administered untimed.

Vocational Interest Career Examination (VOICE). This typically takes 15-20 minutes. The scales include Office Administration (OAD: 20 items); Heavy Construction (HC: 20 items); Electronics (ELE: 20 items); Medical Service (MED: 20 items); Science (SCI: 20 items); Outdoors (OUT: 15 items); Aesthetics (AES: 15 items); Mechanics (MEC: 15 items); Food Service (FS: 15 items); Law Enforcement (LAW: 15 items); Agriculture (AG: 15 items); Mathematics (MTH: 12 items); Audiographics (AUD: 10 items); Teacher/ Counseling (TEA: 10 items); Marksman (MRK: 10 items); Drafting (DFT: 7 items); Craftsman (CFT: 7 items); Automated Data Processing (ADP: 7 items). Coefficient alphas ranged from 0.75 to 0.96 with a median of 0.89.

It was hypothesized that the following scales would be most useful for the selected MOS: HC, MRK, ELE, OAD, and MEC. MEC would be the scale for the Mechanics, OAD would be the scale for the Administrative Specialists, HC and MRK would be scales useful for the Tank Crewman, and OAD and ELE would be useful scales for Radio Teletype Operator. Two additional scales, ADP and MTH, might also be useful in distinguishing the Administrative Specialists from the Radio Teletype Operators.

Personal Opinion Inventory (POI). This typically takes 20-25 minutes. The scales included Conscientiousness (CON: 10 items, from DPQ Unlikely Virtues and PRF Infrequency); Social Potency (SP: 27 items, DPQ Social Potency); Stress Reaction (SR: 36 items, DPQ Stress Reaction); Socialization (SOC: 30 items, from CPI Socialization); Rule Abiding (RA: 9 items, from CPI Socialization); Family Closeness (FC: 7 items, from DPQ Stress Reaction); Effort vs. Luck (LCK: 16 items, from Rotter I/E Scale); Internal Locus of Control (LOC: 29 items, from Rotter I/E Scale). The coefficient alpha for CON was 0.44; excluding this validity scale, the range for coefficients alpha was 0.55 to 0.90 with a median of 0.60.

Owens' Biographical Questionnaire (BQ). This typically takes 20-25 minutes. The scales included Academic Achievement (AA: 8 items); Adjustment (ADJ: 12 items); Athletic Interests (ATH: 2 items); Cultural-Literary (CL: 3 items); Independence (IND: 8 items); Intellectualism (INT: 3 items); Leadership (LEAD: 12 items); Physical Activity/Condition (PA: 15 items); Positive Academic Attitude (PAA: 7 items); Parental Control (PC: 11 items); Parental Closeness (PCLO: 15 items); Sociability-Popularity

(POP: 9 items); Sibling Harmony (SIBH: 5 items); Scientific Orientation (SO: 12 items). Other variables in the BQ requested information about academics and course work. Coefficient alphas ranged from 0.49 to 0.88 with a median of 0.75.

Demographics. The additional variable, of whether a soldier had graduated from high school, was available from Army files.

It was hypothesized that the following temperament and biodata scales would index the motivational and attitudinal variables that could predict attrition for cause: RA, SOC, and SR from the POI and AA from the BQ. Because high school diploma status had proven itself as a predictor of early attrition in much prior research, it also was included, as was the AFQT (cognitive) score.

Criteria

Classification. This variable was the nominal one of which MOS the soldier had begun his/her military service and was available at the time attrition data were collected.

Attrition. For each soldier in the sample, file data were available as of December 31, 1984, indicating whether the soldier was still enlisted or, if not, how the soldier had been discharged. The file data are administrative codes indicating the recorded reasons why the attrition has occurred. Three categories of attrition were developed for this research. These categories, with sample file codes, are: Leave for Good Reason (to attend service academy, medical discharge, hardship); Trainee Discharge Program or TDP; Leave for Bad Reason (drug use, desertion, serious crime). The Trainee Discharge Program refers to a set of administrative procedures which permit a comparatively simple dismissal of soldiers, typically within the first 180 days of service, for "failure to adapt" to the Army. While the behavioral characteristics of this category are imprecise, it appears to refer more to motivational than academic weaknesses which prohibit a soldier from making a satisfactory adjustment to Army life. The best single predictor of such early attrition for males is high school diploma status: High school graduates are much more likely to complete their tours of enlistment.

The numbers of soldiers in each category were as follows:

19A (Tank Crewman): Not Attrit = 2,299; Trainee Discharge Program = 107; Bad Attrit = 73; Good Attrit = 135.

31C (Radio Teletype Operator): Not Attrit = 1,750; Trainee Discharge Program = 141; Bad Attrit = 51; Good Attrit = 47.

63B (Mechanic): Not Attrit = 2,066; Trainee Discharge Program = 33; Bad Attrit = 61; Good Attrit (or Missing) = 37.

71L (Administrative Specialist): Not Attrit = 2,540; Trainee Discharge Program = 121; Bad Attrit = 57; Good Attrit (or Missing) = 77.

The standard procedure is to remove the cases of attrition for good reasons before analysis, which was followed here. The attrition rate was low, ranging from four percent for the Mechanics to ten percent for the Radio Teletype Operators.

Analyses

Data Editing

Predictors. Records were initially checked for consistency of Social Security Number, race, and sex, within person and across inventories. There were several data quality screens for the instruments used here. Details of the procedures can be found in Hough et al. (1984). For the VOICE and the BQ, there was a three-step process to eliminate records which contained too many missing data to yield interpretable scores. There were four steps in the process for the POI, the extra step being the employment of a validity screen via application of the CON scale. Two percent of the VOICE cases and three percent of the BQ cases were deleted because of missing data. Two percent of the POI cases were deleted for the missing data rule, while five percent were dropped because of the CON screen. For item analysis purposes, as well as subsequent analyses, sample sizes varied across scales within these inventories as well as across them.

Criteria. Classification. The criterion here was membership in one of the four MOS, available from the test records.

Criteria. Attrition. As discussed above, file data provided attrition codes so the editing problem was one of matching the PB cases to the file cases.

Descriptive Statistics

For each of the five substantive POI scales, the 19 BQ scales, and the 18 VOICE scales, coefficient alphas were calculated and have been reported above. Means and standard deviations for each scale were calculated for the total sample and for various subgroups of interest as formed by demographic and dependent variables such as MOS, high school diploma status, and attrition category. These descriptive statistics will not be discussed further here.

Inferential Statistics

The uniqueness (U^2) of each predictor scale from the ASVAB was calculated. Uniqueness is the amount of reliable variance of a given variable not predicted by another variable or set of variables. The computational formula is $U^2 = Rxx - R^2$, where U^2 = uniqueness, Rxx = the reliability of the variable of interest, and R^2 = the squared multiple regression when the variable of interest is regressed on some other set of variables. [See Wise and Mitchell (1985) for a more extended treatment of uniqueness.]

The ranges and median values of the uniqueness coefficients for the variables considered here were as follows. For the VOICE, the range was 0.64 to 0.85, with a median of 0.77. For the POI, the range was 0.54 to 0.86, with a median of 0.60. For the BQ, the range was 0.43 to 0.86 with a median of 0.70. Thus, this set of measures is capturing much reliable variance which is not being picked up by the cognitive test battery.

The next step was the computation of a series of analyses of variance with each predictor scale. The major variable was attrition (Not Attrit, Trainee Discharge program, and Bad Attrit) or classification (MOS membership).

The next analyses were in direct reference to the hypotheses stated above. For Hypothesis One, discriminant function analysis was performed using the selected VOICE scales as predictors and MOS memberships as the criterion. For Hypothesis Two, multiple regression analyses were performed, using selected BQ, POI, and VOICE scales to predict attrition status.

Results and Discussion

Hypothesis One stated that the four MOS would differ in the average scores of selected interest scales, as evaluated by the VOICE. Both generalized analyses of variance as well as discriminant analyses showed this to be the case. For the analyses of variance, each of the 18 VOICE scales as well as the AFQT significantly ($p < .01$) discriminated among the four occupations. Discriminant analyses were used with, first, five VOICE scales (HC, MRK, ELE, OAD, MEC) and, second, with two additional scales (ADP, MTH). As displayed in Table 1, both the Administrative Specialists and the Mechanics were fairly well predicted (76% and 69% correct predictions, respectively), somewhat better with the five scales than with the seven scales.

The Tank Crewmen were less well predicted, although adding the two scales to the initial five helped somewhat. The Radio Teletype Operators were predicted least well. The seven scales did somewhat better than the five, but neither group of scales predicted membership in this MOS at much greater than a chance level. One obvious explanation for the difference in predictability among the four MOS lies in the origin of the VOICE instrument. It was designed for Air Force specialties and included occupational scales pertinent to them. Administrative Specialists and Mechanics are common to all military services, hence it is not surprising that these two occupations are well predicted. Tank Crewman and Radio Teletype Operator, on the other hand, are uniquely Army occupations and did not have specific VOICE scales. It should not be surprising that these two occupations were less well predicted.

For Hypothesis Two, the prediction of attrition, the biodata and temperament scales will be considered first. In the generalized analyses of variance, most of the BQ scales and most of the POI scales predicted attrition in the anticipated direction. The nonpredicting scales had more to do with cooperativeness types of variables (e.g., BQ: SIBH, POP; POI: SP) while the predicting scales had more to do with socialization and

achievement traits, (e.g., BQ: AA, LEAD, ADJ, INT; POI: SR, RA, SOC). Most of these scales also correlated with high school diploma status, although not quite as consistently nor as strongly. This probably reflects that many of the variables contributing to a young person's completion of high school are being evaluated by the biodata and temperament measures used here.

Multiple regression analyses of attrition split the criterion into two classes: stay or leave (the latter being both the Trainee Discharge Program and Bad Attrit). Keep in mind that the overall rate of attrition is quite low, at seven percent. This will make prediction difficult. As displayed in Table 2, the traditional predictors of attrition, high school diploma status combined with AFQT, were significantly but mildly related to attrition. The combination of the four hypothesized biodata-temperament scales were also significantly related to attrition, with an adjusted R almost twice the size of that yielded by the traditional predictors. (Other biodata-temperament scales could have been selected but it is likely the results would have been the same.) Combining the two classes of predictors did not improve the prediction in any noticeable way beyond that provided by the biodata-temperament scales. The values of the adjusted R's are relatively small, but recall that the overall rate of attrition to be predicted was also small (only seven percent attrition). Such a severe split on the criterion operates to reduce the expected correlation with other variables. Further, the increase in validity over the traditional predictor of high school diploma status, as provided by the four noncognitive scales, suggests that the latter may provide a useful addition to Army selection and classification procedures.

Table 3 is an expectancy table displaying the predicted attrition rates of soldiers selected on the composite of the four noncognitive scales. As can be seen, using a cutoff standardized score of, say, 25 on this attrition predictor would eliminate two percent of all applicants, but the attrition rate in the cutoff group was predicted to be 28 percent rather than the seven percent over all. A cutoff score of 50, on the other hand, would eliminate half of all these soldiers who would have had an attrition rate only slightly higher (ten percent) than that of the total group. While the choice of a specific cutoff score has many arbitrary aspects to it, it seems clear that a low cutoff score based on a biodata-temperament scale could eliminate a small percentage of soldiers who could be predicted to have an attrition rate significantly higher than average. This result obviously requires replication, both for this group as they continue their Army service, as well as for other, different groups of soldiers.

The second part of Hypothesis Two concerned how well interest measures might predict attrition. Using the same five interest scales included in Hypothesis One, for classification, in addition to the four biodata-temperament scales, yielded adjusted multiple R's ranging from 0.11 to 0.20 across the four MOS. These are displayed in Table 4. The variation in values directly mirrors the base rates of attrition in the MOS to be predicted. The highest R was in the MOS with the highest rate of attrition, Radio Teletype Operator, while the lowest was in the MOS with the lowest rate of attrition, Mechanic. Thus, while it might appear that vocational

Radio Teletype Operator, while the lowest was in the MOS with the lowest rate of attrition, Mechanic. Thus, while it might appear that vocational interest measures could also be useful as predictors of early attrition, this conclusion must be tempered here by the very limiting values of the differing and low attrition rates. That is why no further investigation of specific VOICE scales was undertaken at this point. It would seem that a vocational interest battery specifically tailored to Army jobs (which is part of other Project A research) should be used and evaluated for discriminative efficiency before such measures might be used operationally as predictors of attrition. The case for biodata-temperament measures is much stronger.

In conclusion, this research has demonstrated that noncognitive measures can be effective predictors of two aspects of Army performance, classification and attrition. Classification was better predicted when the occupational interest scale was appropriate to the Army occupation being predicted. While the attrition regression coefficients were relatively low in value, this was probably due, at least in part, to the low base rates of attrition to be predicted. It may be that different analytic methods, such as probit or logit analysis (Aldrich and Nelson, 1984) might better explicate the relationship between attrition and the noncognitive measures used in this research. Follow-on research with this cohort, as its members move through their Army careers, should explore these methods. It is also likely that the rates of attrition will increase as the cohort "ages," operating to improve the chances for accurate prediction.

Acknowledgement

Thanks to Clinton B. Walker, U.S. Army Research Institute, for providing the attrition codes used in this research.

References

Aldrich, J. H., & Nelson, F. D. (1984). Linear probability, logit, and probit models. Beverly Hills; Sage Publications.

Alley, W. E., & Matthews, M. D. (1982). The Vocational Interest Career Examination. Journal of Psychology, 112, 169-193.

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

Gough, H. G. (1975). Manual for the California Psychological Inventory. Palo Alto, CA: Consulting Psychologists Press.

Hough, L. M., Dunnette, M. D., Wing, H., Houston, J., & Peterson, N. G. (1984, August). Covariance analyses of cognitive and noncognitive measures in Army recruits: An initial sample of Preliminary Battery data. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Jackson, D. N. (1967). Personality Research Form Manual. Goshen, NY: Research Psychologists Press.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology Monographs, 56, 347-368.

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

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. Journal of Applied Psychology Monographs, 64, 569-607.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80, (1, Whole No. 609).

Tellegen, A. (1982). Brief Manual for the Differential Personality Questionnaire. Unpublished manuscript, University of Minnesota, Minneapolis.

Wise, L. L., & Mitchell, K. J. (1985, August). Development of an index of maximum validity increment for new predictor measures. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles, CA.

Table 1. Predicting Military Occupation Membership with Selected VOICE Scales

Actual Occupation	Percent of Total Group	Predicted Occupation (5 Scales)* Tank Crewman	Radio Operator	Mechanic	Admin. Spec.
Tank Crewman	27	[44]	27	17	9
Radio Operator	21	14	[26]	7	10
Mechanic	23	31	22	[69]	5
Admin. Specialist	29	12	25	7	[76]
		Predicted Occupation (7 Scales)*			
Tank Crewman	27	[48]	27	19	9
Radio Operator	21	14	[30]	8	10
Mechanic	23	28	18	[67]	6
Admin. Specialist	29	12	25	7	[76]

*The five scales used were Heavy Construction, Marksman, Electronics, Office Administration, and Mechanic. The seven scales were the above five plus Automated Data Processing and Mathematics.

Note: Entries in table are percentages.
Boxed figures are correct predictions.
Columns sum to 100 within rounding errors; rows do not.

Table 2. Predicting Overall Attrition with Traditional and Noncognitive Predictors

Predictors	Used in Prediction Regression		
Traditional			
AFQT	Yes	No	Yes
High School Graduation	Yes	No	Yes
Noncognitive			
Rule Abiding	No	Yes	Yes
Socialization	No	Yes	Yes
Stress Reaction	No	Yes	Yes
Academic Achievement	No	Yes	Yes
Adjusted R	.08	.15	.15
Sample Size	8,352	8,198	7,480

Table 3. Predicted Percent Attrition for Different Cut-off Scores

Predicted Attrition Score* (Standardized) Cut-off	Attrition Rate in Below Cut-off Group	Percent of Total Group in Group Below Cut-off
20	.25	1
25	.28	2
30	.23	3
35	.17	8
40	.15	16
45	.12	31
50	.10	50
55	.08	69
60	.08	85
65	.07	95
70	.07	99
75	.07	100

*Composed of Rule Abiding, Socialization, Stress Reaction, and Academic Achievement Scales.

Table 4. Attrition Status By Military Occupation

	Demographics*		Attrition				
			Numbers of Cases				
Occupation	n	Percent females	Good*** Attrition	Trainee Discharge Program	Bad Attrition	Percent Retention**	Adjusted R***
Tank Crewman	2,614	0	135	107	73	93	.15
Radio Operator	1,989	14	47	141	51	90	.20
Mechanic	2,197	6	37	33	61	96	.11
Admin. Specialist	2,798	48	77	121	57	93	.15

*Occupations were generally 5% Hispanic and 20% Black, except for Administrative Specialist which was 40% Black.

**Good Attrition cases removed before calculation of Retention, at completion of an average of nine months of service.

***Adjusted multiple correlation between binary Attrition Status and four biodata-temperament predictor scales plus five interest scales.

IDENTIFYING OPTIMAL PREDICTOR COMPOSITES AND TESTING FOR GENERALIZABILITY ACROSS JOBS AND PERFORMANCE CONSTRUCTS

Lauress L. Wise
American Institutes for Research

John P. Campbell
Human Resources Research Organization

Norman G. Peterson
Personnel Decisions Research Institute

Presented at the Annual Conference of the
Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

Identifying Optimal Predictor Composites And Testing for Generalizability Across Jobs and Performance Constructs

Industrial psychologists have long been concerned with the problem of matching people to jobs. For a long time, the implicit model in this enterprise was essentially a peg and hole model with job applicants represented by different sizes and shapes of pegs and job openings represented by different sizes and shapes of holes. The goal was to match the pegs to the holes. The primary conclusion drawn from this model, as expressed by Ghiselli (1966), was that different job performance prediction measures and different selection criteria should be developed and validated for different jobs and job environments.

More recently, Schmidt and Hunter (1981) created a paradigm shift when they showed convincingly that a large part of the situational variance in prediction validities is attributable not to differences in job requirements, but to methodological artifacts. Validity generalization is now a household word in Industrial and Organizational psychology. As the term has come to be used, it refers to properties of a distribution of criterion-related validity coefficients generated by using one or more measures of the same construct to predict general job performance within broad families of jobs. The interesting parts of the distribution are its overall mean and the degree to which its variance can be accounted for by statistical artifacts (e.g., criterion unreliability, sampling error) vs. the substantive characteristics of different situations (e.g., different abilities are required by different jobs). Arguments continue as to how to define the appropriate population of coefficients and how large the substantive variance has to be before we should worry about it.

Much of the discussion of validity generalization has focused on the prediction of overall job performance using a measure of general mental ability. Not much attention has been devoted to whether different predictor constructs or different performance components define different populations for validity generalization purposes. From this new perspective, all the pegs are the same shape and the only question is whether they are big enough to fill any particular hole.

In Project A, we have been building a model for both sides of the equation. That is, we are attempting to define the total domain of potentially useful prediction information, describe it in terms of its basic constructs, and then develop representative measures of those constructs (Peterson, Hough, Dunnette, Rosse, Houston, Toquam & Wing, 1987). Similarly for the criterion side, we have tried to define the total domain of job performance, describe it too in terms of basic

factors, and use multiple methods to provide scores on each performance factor (Campbell, Felker, Borman & Rumsey, 1987; Campbell, McHenry & Wise, 1987; Wise, Campbell, McHenry & Hanser, 1986).

Stated simply, our working theory is that performance is not one thing and that the correlations between the major components of performance do not approach the limits of their reliabilities. Similarly, at least some of the basic predictor constructs (latent variables) that account for individual differences at the time of hire are also not highly intercorrelated. As part of Project A, each of these domains has been modeled and measured for a diverse and representative sample of entry-level jobs. Specifically, concurrent validity data have been collected for over 4,000 soldiers in a core sample of nine jobs. With these data we can address such questions such as: How do the validities for each of several predictor constructs generalize across different components of performance? How do the validities of the predictor battery for a particular performance component generalize across jobs? It is to these questions that we now turn.

Method

Data

The data analyzed for this paper included scores for five job performance constructs and twenty-four predictor constructs collected on a sample of several hundred soldiers in each of nine different jobs. Young, Harris, Hoffman & Wise (1987) have described the collection and editing of these data. Table 1 lists the predictor and performance construct scores. Table 2 lists the nine Army jobs and gives the number of soldiers included in the present analyses.

Analyses

Sample covariance matrices, including both the five criteria and the twenty-four predictors, were computed for each of the nine jobs. An overall covariance matrix was computed as the average of these nine matrices, weighting each by the corresponding sample size. This form of pooling was necessary because the criterion measures were comprised of somewhat different items for the different jobs, so that it was not possible to assure fully comparable scaling across jobs.

In the present study, covariances were analyzed as a means of controlling for differences between jobs in heterogeneity with respect to the predictor measures. Initial

selection into each of the nine jobs included an absolute screen on a composite of the subtests from the Armed Services Vocational Aptitude Battery (ASVAB). Different composites and different selection ratios (cutting points) were used for different jobs. For some jobs, the cutoff point was at the population mean, while for others a cutoff as much as .75 s.d. below the population mean was used. In addition to this absolute screen, self-selection and attrition during and after training served to further reduce the heterogeneity of our samples. By including the predictor covariances as an explicit part of our modeling, differences in heterogeneity were accounted for.

The LISREL program (Joreskog & Sorbom, 1981) was used to model the predictor-criterion relationships. This program enables direct statistical tests of the degree to which observed variation in parameter estimates might be due simply to sampling error. The LISREL program also allows separate modeling of the statistical properties, including specifically reliabilities, of the measures analyzed. It is thus possible to eliminate both sampling error and differences in criterion reliability as artifactual sources of variation in predictor-criterion relationships.

In applying LISREL to the present problem, the covariance matrices were divided into three components. The first, the covariances among the predictors, is modeled by the Phi matrix in LISREL. In all of our analyses, the Phi matrix was left unconstrained because differences due to selection were fully anticipated. The second component is the predictor-criterion covariances. In LISREL, these are modeled in the Gamma matrix as regression or structural equations. These equations are used to estimate criterion scores from predictor scores. Most of our analyses consisted of testing possible constraints on this matrix (i.e., constancies across criterion constructs or across job samples). The final component into which the covariance matrices were divided was the Psi matrix. PSI contains the covariances among the unique/error portion of the criterion measures. In LISREL, the observed criterion covariances are modeled as the sum of the covariances among estimated criterion scores and the covariances among the error/unique components of the criterion variables.

The first step in our analyses was to reduce the number of predictor scores included in the model. This was done to simplify our representation of predictor-criterion relationships and to make the subsequent structural equations more stable. The approach used was to successively eliminate predictors and then examine whether all of the predictor-criterion covariances could be adequately reproduced without including the eliminated predictors in any of the structural (regression) equations. In such a case, the predictors in

question could be dropped without loss of any predictive information.

The second step was to test for criterion equivalence. If two or more criterion constructs shared a common set of relationships to the predictors, then further reduction of the criterion space would be possible. For each pair of criterion constructs, a model was fit to the combined covariance matrix in which the regression coefficients for each predictor were constrained to be the same in both criterion equations. If this model was not rejected by the data, then the criteria could be combined without loss of information concerning predictor-criterion relationships.

The third and final step in our analyses was to test for equivalence in the prediction equations for different jobs. For each distinct performance construct, a model with a constant prediction equation across all nine jobs was tested.

Results

Table 3 shows standardized regression coefficients for predicting each criterion construct from the entire set of predictor constructs. This was our starting point for reducing the number of different predictors considered. We examined the significance of each of these coefficients, using the t statistics provided by LISREL (testing for difference from zero). We eliminated those predictors that did not have a significant coefficient ($t > 2.0$) for any of the criteria. This resulted in a model which did not quite fit the overall covariance matrix, so we put those predictors with the largest modification indices back into the equations. (Note that each predictor was either in all of the equations or none at this stage.) In the end, five predictors were eliminated. Each of the remaining nineteen predictors had a significant loading on at least one of the criteria. Table 4 shows the reduced structural equations and gives fit statistics for this reduced model.

Table 5 shows the results of the tests for criterion equivalence. In all cases, separate prediction equations were indicated. The Core Technical Proficiency and General Soldiering Proficiency constructs were the most similar. The three primary differences between the equations for these two constructs, as seen in Table 5, were: (1) the distinctly greater significance of the Combat Interest measure for predicting General Soldiering Proficiency; (2) the somewhat greater significance of spatial and quantitative skills for General Soldiering Proficiency and (3) the somewhat greater significance of verbal skills for Core Technical Proficiency. It also was the case that high scores on the Physical Con-

ditioning predictor were related to lower Core Technical Proficiency scores but not to lower General Soldiering Proficiency scores. The differences between the equations for the other criterion equations were all in expected directions and fully consistent with the general findings reported by McHenry et al. (1987).

Further analyses were conducted to identify optimal sets of predictors for each of the criterion constructs. An initial model, in which only cognitive and perceptual tests were used in predicting proficiency and only interest, temperament, and biographical measures were used in predicting the motivational constructs did not fit the data adequately. A number of iterations were performed with changes based on the data. Table 6 shows the predictors and their standardized coefficients that were judged to best fit the data. Given some reliance on empirical results in identifying this model, the significance level should not be overinterpreted.

Table 7 shows the results of tests for equivalent prediction equations across jobs for each criterion construct. In these analyses, a reduced set of predictors was used for each performance construct. This was done partly because multisample runs can otherwise be inordinately expensive and partly in an effort to achieve some semblance of parsimony.

For the three "will do" performance constructs (Effort and Leadership, Maintaining Personal Discipline, and Physical Fitness and Military Bearing), the hypothesis that one equation fits all jobs could not be rejected from the available data (shown by Chi-square statistics with p values greater than .05.) For the General Soldiering Proficiency construct, the p value fell between .01 and .05, suggesting at most very modest differences in the prediction equations across jobs.

For Core Technical Proficiency, however, the common prediction equation model was strongly rejected. Table 8 shows the separate prediction equations estimated for each job. Table 9 shows chi-square fit statistics and p values for each pair of jobs considered by themselves. For some pairs of jobs, such as the combat jobs, the optimal prediction equations were not significantly different. For other jobs, however quite significant differences were found. The largest difference was between Vehicle Mechanics and Administrative Specialists.

The results presented in Tables 8 and 9 suggest that there are significant differences between the requirements of mechanical/technical jobs, clerical/administrative jobs, and combat jobs. It is not clear whether there are differences in job requirements within each of these three major job types, but our data suggest that this might be the case.

DISCUSSION AND SUMMARY

The first general result from these analyses is that there are different components of job performance, even within entry-level positions, that show different patterns of relationships with potential predictors measures. The predictors of job proficiency were, for the most part, quite distinct from predictors of effort and leadership, avoidance of disciplinary problems, and physical fitness/military bearing. These results generally supported our perspective that job performance is, indeed, multidimensional and not just one thing. One consequence of this result is that the assessment of overall job effectiveness necessarily involves policy decisions regarding the relative importance of the different components of job performance in a particular setting. Project A staff are now in the process of collecting such judgments for each of the jobs included in our sample.

The second general result is that <u>different</u> mixes of skills, interests, temperament and background must be used to obtain optimal prediction of technical proficiency in different jobs. These results are particularly important for an organization like the Army, which must select and train untrained individuals for many different jobs. Individual differences in job knowledge prior to training would not necessarily be related to differences in job proficiency after training, yet we still find significant differentiation in predictors of post-training job performance. These results suggest that the test studied would be useful for classifying new Army recruits into jobs that are best suited to their abilities, temperament, and interests.

The results of the present analyses undoubtedly understate differences between jobs in a number of ways. First, the common Army experience and environment shared by the soldiers who served as subjects in Project A probably increases similarities in job requirements. To succeed in the Army, all soldiers must pass basic training and advanced technical training. All soldiers must learn and adhere to Army customs and tradition. And, to at least some degree, all soldiers must subscribe to Army values. These similarities attenuated the differences in job requirements that we uncovered in our analyses. Second, all soldiers share a number of responsibilities, regardless of their assigned MOS. These shared responsibilities are reflected in the job performance constructs; General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing. These performance constructs were intended to capture job performance components that are common to all soldiers. Thus, we were not surprised to discover that the optimal

predictors of these performance constructs were the same across all jobs -- even though it reduced our power to differentiate between jobs on the basis of skill and trait requirements. Third, we have studied only entry-level positions within each of these jobs. Increasing differentiation between jobs seems likely as incumbents graduate to more skilled positions. This possibility is being addressed in continuing Project A activities aimed at assessing the performance of more experienced job incumbents.

One final caveat is a reminder that, with the exception of the ASVAB scores, the predictor and criterion data analyzed here were collected concurrently, differences in predictive relationships may have been either worn down by common Army experience or accentuated through differences in training and on-the-job experiences. We are now engaged in the longitudinal phase of Project A which will allow us to assess the range of opportunities for matching individuals to jobs even more conclusively than was possible in the present analyses.

REFERENCES

Campbell, C.G., Borman, W.C., Felker, D.C., Ford P., Park, M.D., Pulakos, E.C., Riegelhaupt, B.J., & Rumsey, M.G. (1987, April). Development of project A job performance measures. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Campbell, J.P., McHenry, J.J., & Wise, L.L. (1987, April). Analysis of criterion measures: The modeling of performance. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Ghiselli, E.E. (1966). The validity of occupational aptitude test. New York: Wiley.

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

McHenry, J.J., Hough, L.M., Toquam, J.L., Hanson, J.A., & Ashworth, S. (1987, April). Project A validity results: The relationship between predictor and criterion domains. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Peterson, N.G., Hough, L.M., Dunnette, M.D., Rosse, R.L., Houston, J.S., Toquam, J.L., & Wing, H. (1987, April). Identification of predictor constructs and development of new selection/classification tests. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Schmidt, F.L., & Hunter, J.E. (1981). Employment testing: Old theories and new research findings. American Psychologies, 36, 1128-1137.

Wise, L.L., Campbell, J.P., McHenry, J.J., & Hanser, L.R. (1986, August). A latent structure model of job performance factors. Paper presented at the 92nd Annual convention of the American Psychological Association, Washington, D.C.

Young, Y.Y., Harris, J.H., Hoffman, G.R., & Wise, L.L. (1987, April). Large scale data collection and data base preparation. Paper presented at the Second Annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Table 1
Predictor and Job Performance Constructs

Predictor Constructs	
Type of Measure	Construct Name
Armed Services Voc. Aptitude (ASVAB)	Quantitative
	Speed
	Technical
	Verbal
Project A: Cognitive	Spatial
Project A: Percept. and Psychomotor	Complex Perceptual Accuracy
	Complex Perceptual Speed
	Numerical Speed and Accuracy
	Psychomotor Ability
	Simple Reaction Accuracy
	Simple Reaction Speed
Project A: Temperament	Adjustment
	Dependability
	Physical Conditioning
	Achievement Orientation
Project A: Interest	Audio/Visual Interest
	Combat Interest
	Food Service Interest
	Protective Service Interest
	Skilled Technical Interest
	Structural/Machines Interest
Project A: Job Orientation	Job Autonomy
	Organizational/Coworker Support
	Routine Work

Performance Constructs	
Type of Measure	Construct Name
Hands-on and Written Tests	Core Technical Proficiency
	General Soldiering Proficiency
Administrative Measures and Ratings	Effort and Leadership
	Personal Discipline
	Physical Fitness and Military Bearing

Table 2
The Number of Incumbents in the Nine Army Enlisted Jobs Studied

Enlisted Job	Number of Incumbents
Infantryman	491
Cannon Crewmember	464
Armor Crewmember	394
Single Channel Radio Operator	289
Light Wheel Vehicle Mechanic	478
Motor Transport Operator	507
Administrative Specialist	427
Medical Specialist	392
Military Police	597
Total	4039

Table 3

Standardized Regression Coefficients for Each Criterion Against All Predictors (Based on Pooled Covariance Matrix, n=4039)

Predictor	Core Technical Proficiency	General Soldiering Proficiency	Effort and Leadership	Personal Discipline	Physical Fitness/ Military Bearing
Quantitative	.097	.130	.012	.063	.023
Speed	.020	-.011	.062	.032	.088
Technical	.103	.141	.155	.082	-.038
Verbal	.116	.098	-.080	-.029	-.106
Spactial	.196	.279	.014	-.005	-.021
Complex Perc. Accy.	.085	.112	.046	.026	.015
Complex Perc. Speed	.032	.039	.052	.033	.032
Num. Speed/Accy.	.032	.020	.016	-.028	-.026
Psychomotor	.003	.047	.020	-.029	-.011
Simple Reaction Accy.	.012	-.004	-.020	.011	-.036
Simple Reaction Speed	.026	.028	-.014	-.019	.044
Adjustment	-.004	.000	-.004	.001	.025
Dependability	.127	.128	.119	.314	.099
Physical Condition	-.053	-.007	.008	-.054	.248
Achiev. Orient.	-.003	-.045	.221	.040	.131
Audio/Visual Interest	-.054	-.008	-.026	-.025	.038
Combat Interest	.103	.167	.117	.017	.042
Food Service Interest	-.050	-.036	-.060	-.042	-.021
Prot. Service Interest	-.009	.003	.011	-.033	-.051
Skilled Tech. Interest	-.010	-.020	-.031	-.009	.009
Structural/Machine Int.	.054	-.007	-.011	-.026	-.052
Job Autonomy	.014	-.007	.000	-.042	-.046
Job Support	.034	.023	-.020	-.020	.017
Routine Work	-.023	-.039	-.037	-.015	-.010
R-Squared	.223	.305	.146	.114	.160

NOTE: Values are maximum likelihood estimates which may differ slightly from OLS estimates presented elsewhere. Also, no attempt was made to estimate parameters for the unselected population.

Table 4

Standardized Regression Coefficients for Each Criterion Against Reduced Predictor Set (Based on Pooled Covariance Matrix, n=4039)

Predictor	Core Technical Proficiency	General Soldiering Proficiency	Effort and Leadership	Personal Discipline	Physical Fitness/ Military Bearing
Quantitative	.088	.119	.002	.058	.018
Speed	.025	-.006	.065	.033	.091
Technical	.102	.142	.159	.084	-.039
Verbal	.123	.106	-.075	-.027	-.101
Spatial	.205	.293	.034	.007	-.012
Complex Perc Accy	.070	.093	.022	.011	.000
Complex Perc Speed	.	.	.	.	.
Num. Speed/Accy	.039	.027	.024	-.023	-.019
Psychomotor	.007	.051	.025	-.025	-.007
Simple Reaction Accy	.012	-.004	-.020	.011	-.037
Simple Reaction Speed	.033	.036	-.004	-.012	.052
Adjustment	.	.	.	.	.
Dependability	.126	.123	.108	.308	.100
Physical Condition	-.053	-.006	.011	-.053	.249
Achiev. Orient.	.005	-.034	.220	.039	.153
Audio/Visual Interest	-.054	-.014	-.041	-.030	.044
Combat Interest	.102	.167	.118	.017	.042
Food Service Interest	-.058	-.047	-.069	-.045	-.023
Prot. Service Interest	-.006	.005	.008	-.034	-.048
Skilled Tech Interest	.	.	.	.	.
Structural/Machine Int.	.051	-.014	-.019	-.029	-.052
Job Autonomy	.021	-.002	-.004	-.047	-.045
Job Support	.	.	.	.	.
Routine Work	.	.	.	.	.
R-Squared	.221	.303	.143	.113	.159

NOTE: Chi-Square = 35.07, df=25, p=.09. A "." denotes that the predictor was not included in the regression model.

Table 5

Chi-Square Values for Tests for Common Regression Equations for Each Pair of Criteria

	Core Technical Proficiency	General Soldiering Proficiency	Effort and Leadership	Personal Discipline
General Soldiering Proficiency	84.6			
Effort and Leadership	374.5	416.6		
Personal Discipline	513.1	754.2	566.6	
Physical Fitness/Mil. Bearing	973.2	1230.9	559.6	542.0

NOTE: All Chi-squares had P values less than .001, indicating rejection of the common predictor equation model.

Table 6

Standardized Regression Coefficients Using Selected Predictors for Each Criterion (Based on Pooled Covariance Matrix, n=4039)

Predictor	Core Technical Proficiency	General Soldiering Proficiency	Effort and Leadership	Personal Discipline	Physical Fitness/ Military Bearing
Quantative	.084	.114	.	.055	.
Speed	.027	.	.045	.	.077
Technical	.101	.137	.140	.059	-.047
Verbeal	.132	.113	-.061	.	-.092
Spatial	.212	.295	.034	.	.
Complex Perc. Accy.	.064	.086	.	.	.
Num. Speed/Accy.	.049	.035	.038	.	.
Psychomotor	.	.054	.041	.	.
Simple Reaction Accy.	.	.	.	.	-.031
Simple Reaction Speed	.	.	.	.	.054
Dependability	.127	.124	.112	.314	.100
Physical Condition	-.049	.	.	-.062	.245
Achiev. Orient.	.	-.040	.222	.041	.156
Audio/Visual Interest	-.046	.	-.043	-.034	.039
Combat Interest	.097	.158	.108	.	.036
Food Service Interest	-.063	-.051	-.060	-.038	.
Prot. Service Interest	.	.	.	-.040	-.053
Structural/Machine Int.	.059	.	.	.	-.047
Job Autonomy	.	.	.	-.046	-.041
R-Squared	.220	.302	.140	.110	.155

NOTE: Chi-Square = 32.86, df=38, p=.71. A "." denotes that the predictor was not included in the regression model for that criterion.

Table 7
Test for Common Prediction Equations Across Nine Jobs

CRITERION	CHI-SQUARE	DF	P
Core Technical Proficiency	220.5	65	.000
General Soldiering Proficiency	80.7	57	.02
Effort and Leadership	69.7	57	.12
Personal Discipline	91.3	73	.07
Physical Fitness/Mil. Bearing	111.3	89	.06

Table 8

Standardized Regression Coefficients for Predicting Core Technical Proficiency Overall and for Each

Predictor	11B: Infantry	13B: Cannon Crew	19E: Armor Crew	31C: Radio Oper.	63B: Vehicle Mechanic	64C: Truck Driver	71L: Admin. Spec.	91A: Medic Spec.	95B: Military Police	All
ASVAB										
Quantitative	.101	.030	.038	.247	-.018	.058	.301	.053	.106	.096
Technical	.105	.002	.164	.133	.436	.284	-.179	.088	.036	.101
Verbal	.162	.067	.100	.127	-.030	-.021	.088	.245	.110	.100
Project A										
Spatial	.258	.218	.209	.101	.080	.163	.250	.225	.177	.197
Complex Prec Accy	.045	.052	.101	.095	-.028	.093	.149	.019	.078	.063
Dependability	.066	.056	.080	.118	.071	.079	.144	.226	.116	.099
Combat Interest	.164	.210	.156	.034	.157	.062	-.062	.106	-.010	.097
Food Serv. Int.	-.086	-.101	-.052	.060	-.057	-.041	-.048	-.018	-.067	-.057
R-Squared										
Separate Pred.	.350	.163	.291	.224	.349	.225	.300	.291	.136	n/a
Combined Pred.	.336	.126	.283	.184	.270	.197	.186	.253	.117	.214

Table 9
Validity Generalization for Core Technical Proficiency Across all Batch A MOS
Chi-Square Values/Probability level

	11B: Infantry	13B: Cannon Crewman	19E: Armor Crewman	31C: Radio Operator	63B: Vehicle Mechanic	64C: Truck Driver	71L: Admin Specialist	91A: Medical Specialist	95B: Military Police
13B	19.82 0.048								
19E	16.72 0.116	11.61 0.396							
31C	19.39 0.054	19.11 0.059	20.78 0.036						
63B	33.07 0.001	31.38 0.001	21.43 0.029	45.82 0.000					
64C	28.88 0.005	17.21 0.102	6.61 0.830	20.93 0.034	19.96 0.046				
71L	55.95 0.000	49.59 0.000	56.93 0.000	29.71 0.002	117.48 0.000	66.60 0.000			
91A	19.87 0.047	34.48 0.000	28.26 0.003	19.74 0.049	50.74 0.000	42.80 0.000	59.33 0.000		
95B	17.04 0.107	20.80 0.035	24.26 0.012	16.73 0.116	57.38 0.000	26.21 0.006	32.99 0.001	27.40 0.004	

LARGE-SCALE DATA COLLECTION AND DATA BASE PREPARATION

Winnie Y. Young
American Institutes for Research

Janis S. Houston
Personnel Decisions Research Institute

James H. Harris
R. Gene Hoffman
Human Resources Research Organization

Lauress L. Wise
American Institutes for Research

Presented at the Annual Conference of the
Society for Industrial and Organizational Psychology

Atlanta, Georgia

April 1987

Large Scale Data Collection and Data Base Preparation

In the summer and fall of 1985, data were collected for the concurrent validation phase of the Army's Project A. An extensive array of predictor and criterion measures were administered to approximately 9,500 entry-level soldiers and ratings of these soldiers' performance were also obtained from approximately 7,000 supervisors. The original Project A Research Plan specified a concurrent validation target sample size of 600 job incumbents for each of the 19 jobs or Military Occupation Specialties (MOS), using procedures that had been tried out and refined during the predictor and criterion field tests. The Research Plan further specified that data would be collected at 13 separate sites (Army Posts) in the United States and two in Europe. Individual sites were selected on a basis that maximized the probability of obtaining the target sample sizes without exceeding the project budget.

The logistics involved in such a large-scale data collection effort are fairly complicated and the sheer volume and complexity of the resultant data base presents a challenge to ensure that the data available for analysis will be the highest quality possible. This paper describes our attempt to meet the logistical demands of the concurrent validation data collection and the procedures we used to assemble and edit the resultant data base.

Data Collection

Sampling Plan

The general sampling plan was to use the Army's World-Wide Locator System to identify all the first-term enlisted personnel in the 19 target MOS at each of the 15 selected sites who entered the Army between 1 July 1983 and 30 July 1984. The intent was to represent as many Army "units" as possible while preserving enough cases within units to provide a "within rater" variance estimate for the supervisor and peer ratings of job performance. A two-step sampling procedure was followed.

1. For each site, identify a subset of the 19 target MOS from which it would be possible to draw a large enough sample. That is, given the entry date "window" and given that only 50-70 percent of the

people on any list of potential subjects could actually be found and tested, what MOS are large enough to warrant sampling them at that site?

2. For each MOS in the subset identified above, identify the smallest "unit" from which 6-10 people can be drawn. Ideally, we wanted to sample 4 to 6 units from each site and 6 to 12 people from each unit. For the total concurrent sample this would provide enough units to average out or account for differential training effects and leadership climates, while still providing sufficient degrees of freedom for investigating within-group effects such as rater differences in performance appraisals and in work environment descriptions.

This procedure yielded a rather elaborate matrix of all MOS by site by unit combinations that could reasonably be sampled. From this, a specific sampling plan was prepared for each site that represented the most efficient way possible by obtaining our target across sites of 600 soldiers in each of the 19 MOS.

For a few MOS, there were fewer than 600 soldiers available across all 15 sites with the appropriate accession dates. The decision was made to slightly over-sample the remaining MOS, so our total target sample was still approximately 11,400.

Preparation for Data Collection

Obtaining support. Work began over a year in advance of the actual data collection to obtain the support necessary to reach our target sample. Troop Support Requests (TSR) had to be submitted far in advance, detailing the purpose of the data collection, the schedule of events, the locations, the number of hours required of each soldier, and the complete personnel, classroom, and equipment requirements. After the TSR were submitted, senior project staff met with the Chief Executive Officers (four star generals) of the organizations providing support. Numerous briefings were conducted at various points down the chain of commands, culminating in a two-day meeting with the Point of Contact (POC) assigned to this effort at each site, six months prior to data collection at that site. From this point on, we coordinated primarily with the POC, who was responsible for providing the required troops to be tested, test scorers and other support personnel, equipment, classrooms, etc. This sequence of activities could be summarized as twelve months of planning, briefing, coordinating, cajoling, visiting and monitoring.

Training data collection teams. In order to cover 15 sites in a relatively short period of time, where each site required four to eight weeks of testing, several data collection teams had to be assembled and trained.

Each data collection team was composed of a Test Site Manager and six or seven team members who were responsible for predictor and criterion administration. Test Site Managers were selected from regular project staff who had participated heavily in the field tests. The remaining team members were made up of a combination of regular project staff and individuals (e.g., graduate students) specifically recruited for the data collection effort. This team was assisted on-site by eight Non-commissioned Officer (NCO) scorers (for the Hands-On tests), one company-grade officer POC, and up to five NCO support personnel.

The data collection teams were given three days of training at a central location. During this period, Project A was explained in detail, including its operational and scientific objectives. After the logistics of how the team would operate (transportation, meals, etc.) were discussed, the procedures for data entry from the field to the computer data base were explained in some detail. Emphasis was placed on how to reduce entry errors by ensuring careful recording of responses and correct identification of answer sheets and computer diskettes.

Next, each predictor and criterion measure was examined and explained. The trainees took each predictor test in its entirety, and worked through samples of each criterion measure. Considerable time was spent on the nature of the performance rating scales, rating errors, rater training, and the procedures to be used for administering the ratings. All predictor and criterion administration manuals, which had been prepared in advance, were studied and reviewed, role playing exercises were conducted, and hands-on instruction for maintenance of the computerized test equipment was given.

The intent was that by the end of the three-day session each team member would (a) be thoroughly familiar with all predictor tests and performance measures, (b) understand the goals of the data collection and the procedures for obtaining these goals, (c) have practiced administering the instruments and received feedback, and (d) be committed to making the data collection as error-free as possible.

As noted above, eight NCO scorers were required for Hands-On test scoring. Training for these scorers took place on site over one full day and consisted of (a) a thorough

briefing on Project A, (b) an opportunity to take the tests themselves, (c) a check-out of the specified equipment, and (d) multiple practice trials in scoring each task, with feedback from the project staff.

Data Collection Procedures

Each soldier was tested for a total of either 16 or 8 hours, depending on whether he/she was in a "Batch A" MOS (for which we had MOS-specific criterion measures) or a "Batch Z" MOS (for which we did not). Some of the testing could only be done in fairly small groups because of the equipment required, e.g., Hands-On criterion tests and computerized predictor tests. To accommodate this restriction and still process the maximum possible number of soldiers each day, the predicator and criterion measures were arranged in four-hour testing blocks, each conducted in a separate location. A group of soldiers could then be separated randomly into subgroups and the subgroups rated through the separate testing blocks. The measures administered in each testing block are shown in Figure 1.

Data Base Preparation

Description of Data Base

A total of 9,430 entry-level soldiers in 19 MOS were tested during the concurrent validation data collection. This represents approximately 83% if the total target sample of 11,400. Figure 2 presents a breakdown of this sample by site and by MOS. The MOS are grouped by "batch". Recall that for nine MOS, designated Batch A, an extensive array of criterion measures was developed and administered, including a number of MOS-specific measures. For the remaining 10 MOS, designated Batch Z, an abbreviated set of criterion measures was used.

All of 19 MOS received the same set of predictors. A complete listing of these predictor and criterion measures appears below:

A. Predictors:

- Paper-and pencil tests: six cognitive ability tests
- Computer battery: 10 perceptual/psychomotor tests

B. Criteria:

- Hands-on tests: observation and scoring of performance on 14-17 carefully sampled job tasks (Batch A only)
- Job knowledge tests: written tests of facts and procedures for 30 carefully sampled job tasks (Batch A only)
- School knowledge tests: written tests of facts and procedures taught during training for the MOS
- Ratings of performance by peers and supervisors on several sets of rating scales, including:
  - 11 Army-Wide Behavior Summary Scales
  - 8 to 13 MOS-Specific Behavior Summary Scales (Batch A only)
  - 15 Job Task Rating Scales (Batch A only)
  - 11 Common Task Rating Scales (Batch Z only)
  - 40 Combat Performance Prediction Scales
- Self-report of administrative and personnel records, including:
  - letters and commendations
  - Physical Readiness Test Score
  - Marksmanship Score
  - disciplinary actions (Articles 15, Flag Actions)
- Job History: 'how often' and 'last time' 30 sampled job tasks were performed (Batch A only)
- Work Environment: ratings of 99 items concerning situation at work

Table 1 illustrates the sheer volume of data that were collected.

Editing and Preparation of Data Files

One of the first steps in dealing with so many different instruments, collected at different times at different testing stations, was to match up all of these pieces of information using the common identifier, in our case the Social Security Number (SSN). Before any attempt was made to edit each individual file for merging, a Link file was created. Basically, the Link file consisted of the SSN for each soldier for whom we had any data and a flag for each data source. The idea was to build a relatively manageable file and to resolve all the problems concerning the identifiers before merging.

We found that, although soldiers could reliably write their SSN on a piece of paper, they did not always 'grid' them correctly on our machine scannable forms. In general, we found about 5 percent SSN errors in our sample. There is no simple way to identify the erroneous digit(s), so a great deal of time was spent matching unmatched records by hand. In addition to editing the SSN in our Link file, we also spent some time editing a selective set of demographic variables, including sex and race.

It was not sufficient to verify that variables such as sex and race were within range. In order to identify "errors" on theses variables, we merged two other Army data sources with our data base and compared all three sources for discrepancies. In the case of sex codes, we frequently inspected the soldiers' first names to resolve differences. In the case of race variables, we used a two-out-of-three majority rule.

After the initial editing of basic identifiers and demographic variables was complete and before different pieces of data were merged and ready for analysis, there were many issues that needed to be resolved. These issues included: random responding, missing data, different testing conditions and equipment differences. The decisions that were made regarding how to deal with each issue have subsequently proved to be extremely important to the analyses that were performed.

Random Responding. For multiple choice tests, it is not uncommon for some responders to randomly mark on the answer sheets. This is particularly true when there is no real incentive for taking the test to begin with, unlike tests such as the SAT or GRE. Three different procedures were developed to identify random responding. The first method was by reviewing the test administrators' "problem" logs. At each site, the test administrator was instructed to write down any unusual situations during testing, e.g., it was obvious that someone was not taking the test seriously. Each entry on these logs was entered on a computer file with SSN and a code for that particular "problem". These problems included things such as responded randomly, refused to cooperate and fell asleep etc. Scores were not computed for tests we were told had been completed at random.

The second method used to detect random responding was to score the eight items that were developed for this purpose and embedded in one of the predictors. These items had an extremely obvious "right" answer. For example:

The branch of the service that deals most with airplanes is the:

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

If a soldier got three or more of these eight items wrong, he/she was assumed to be responding randomly and scores were not computed for that instrument.

The third method used was a random response index. For the written tests, a random response index was defined as the correlation between the item score (1 for correct and 0 for incorrect) and item difficulty (expressed as the proportion of subjects who answered the item correctly). For most soldiers this correlation was positive since there was a tendency to get the easier items correct and miss the more difficult items. In a few cases this correlation was essentially zero, suggesting random responding. For these subjects, all of their responses for that particular instrument were set to missing.

For the performance rating scale data, we screened for unreliable raters. We constructed reliability indices for each rater by comparing their ratings with the average of all other raters' ratings of the same soldiers on the same scales. Both mean difference and correlational indices were used in identifying "outliers" among the raters.

Missing Data. No matter how carefully any data collection is planned and monitored, some amount of missing data is inevitable for various reasons. Some of these reasons for missing data in our data set are shown in Figure 3.

One option for dealing with missing data is to delete all of the records for any soldier who had any missing data. This was not an acceptable procedure for our data set. Table 2 shows the number of Batch A soldiers with different patterns of complete and missing data across the four main performance measurement methods: School Knowledge Test (SK), Job Knowledge Test (JK), Hands-on Tests (HO) and Ratings (RA). Fewer than 15% of the cases in the entire sample have complete data for all four methods. If the ratings data are set aside, there are still fewer than 25% of the subjects with complete hands-on, job knowledge, and school knowledge data. Ignoring the hands-on data still leaves only about 42% of the subjects with complete data on the remaining measures. Even if one was willing to conclude that the sample of soldiers with complete data is representative of the target population, the sheer loss of statistical power associated with the reduced sample size would be unacceptable.

The processing of missing data was approached in two stages. In the first stage, we focused on one instrument at a time and dealt with only those subjects who were missing a small amount of data on the instrument under consideration. In the second stage, we formulated procedures for dealing with subjects who were missing a high percentage or all of the data on a given instrument.

Stage I: Missing Data within Each Instrument. We examined the distribution of the missing data for each instrument and found that most were bimodal. Most soldiers had only a small number of missing items or scales but a small number had all or nearly all elements missing. For cases with minimal missing data (usually a ten percent limit was used), we filled in missing values so as to be able to compute overall performance scores. (The procedure used for imputing values is discussed at the end of this section.) For cases with larger amounts of missing data, we did not attempt to compute any scores for the instrument in question. In general, we sought to retain 90 - 95% of the soldiers tested in each MOS, but to eliminate cases with more than 10% missing elements.

Hands-on measures have a different pattern of missing data that warrant a more detailed discussion here. First of all, for several MOS, the hands-on scoring differed for different equipment. In order to achieve comparable scores across these equipment differences, wesplit the examines into separate "tracks" corresponding to the different variations in equipment. For Military Police (95B), for example, females use and were tested on a .38 caliber hand gun while males use and were tested on a .45 caliber hand gun. We found minimal differences between track samples on those tasks that were scored the same, so we achieved comparable scoring by standardizing scores computed from tracked tasks separately for each track sample. Scores for each track were standardized to have a mean and standard deviation that matched the original overall mean for the score in question.

We also checked for anomalies in the Hands-On data such as outliers in quantitative scores, incompatible pass-fail patterns of scores, incorrect coding of soldier ID numbers, and incorrect coding of tracked tests for soldiers. This examination produced 1200 queries. In about 3/4 of the cases, we were able to reach some resolution which provided scores for missing data points or corrected scores for incompatible scores by retrieving the original scoresheets. For example, one common problem was for scorers to write a note that a soldier could not do any of the steps in a task, and then not mark the "NO-GO" (fail) columns for those tasks on the scoresheet. These "missing" scores were changed to "NO-GO" scores.

By far, the greatest reason for non-equivalence in data sets across soldiers was related to equipment. In six of the nine MOS, anticipated variations in equipment necessitated the preparation of tracked versions of tests. Two of these tracks resulted in whole tasks not being administered to a large number of soldiers; an additional 14 tasks were tracked with either completely separate versions of the test, or branching within steps of the test. In addition, unanticipated variations in equipment required Hands-On test managers (project staff) to modify tests in three MOS on site by specifying steps that should not be scored on existing tests Seven tasks had such "last-minute" tracks. On an MOS by MOS basis, rules were established, consistent with Army Doctrine, for equating performance scores across these equipment variations. Discounting these planned and unplanned tracks, most of the remaining cases of incomplete data resulted from unavailable or faulty equipment. Thirteen percent of the task tests could not be administered because of unavailable equipment and two percent could not be scored due to faulty equipment.

The remaining cases of missing Hands-On data were due to a variety of circumstances, most of them unavoidable, such as soldiers who had physical handicaps that prohibited them from performing certain activities, injuries, illness and competing demands for the soldiers time.

After dropping cases with too much missing data or with random responses, we imputed values for the remaining missing data so that summary scores could be computed. The option that we used to fill in missing values was a procedure that had been developed for the National Center for Education Statistics (now the Center for Education Statistics) known as PROC IMPUTE.[1] Several features of PROC IMPUTE made it preferable to other readily available options for filling in the missing values.

---

[1]Wise, L.L. & McLaughlin, D.H. (1980). Guidebook for the imputation of missing data. Palo Alto, CA: American Institutes for Research.

PROC IMPUTE uses regression equations to predict missing values and also adds a random variable with variance equal to the error of estimate for predicting the missing value such that the imputed values are not highly correlated with values on other nonimputed values.

PROC IMPUTE was used in all instances except one. For the written tests, a distinction was made between internal omits (prior to the last item answered) and items that were not reached (omits after the last item answered). For internal omits, we assumed that the examinee did not know the answer and substituted a score equal to the guessing rate (e.g., .2 for a 5 option item). If the actual proportion passing the item was lower than the guessing rate, the proportion passing was used instead. We made no assumptions regarding items not reached since the examinee may not have had time to demonstrate knowledge of the item. Not reached items were imputed with PROC IMPUTE, as were all missing hands-on steps and rating scales.

Stage II: Missing Instruments. After cases were dropped or missing values were filled in on an instrument-by-instrument basis, we were ready to compute overall performance scores that combined information from the different measurement methods. The decision at this stage was whether to estimate individual scores if only partial data were available for the individual. We decided on a 50% rule. An examinee had to have data on at least half of the instruments going into a particular performance construct before we would estimate a score on that construct. Where 50% or fewer of the pieces were missing, PROC IMPUTE was again used to fill in the missing pieces.

Table 3 shows the number of soldiers in each MOS who had missing values for each instrument after the completion of the Stage I imputations and screening. In most instances, the number of missing cases was quite small (1 or 2%). The chief exceptions were two of the administrative measures. (Administrative measures were not included in Stage I imputations because they do not include a large number of component parts.) Physical Readiness Test scores were missing for 10 to 15% of the examines. In most instances, peer and supervisor ratings of physical fitness were available for these same examines. Similarly, Promotion Rate Deviation scores were missing for a significant number of cases (15%). This was primarily due to problems in retrieving Accession File information needed to compute time-in-service.

Summary

Collecting data from 10,000 soldiers in 15 locations over six months is a difficult task, one that requires careful planning, attention to detail, an ability to adapt, a fondness for crisis management, and a special relationship with the telephone. For anyone planning an effort of like grandeur (or even grander), a few lessons learned from some of the survivors seems appropriate.

Planning. Start as early as possible (18 months before collecting data) to identify the support you will need, to include personnel, equipment, facilities, and time requirements. Once you know what you need and when you need it, schedule a series of briefings. Start at the top with the Chief Executive Officers of the organizations who provide the support and work your way through a series of briefings until you reach the local POC responsible for seeing that you get what you need when you need it. Be prepared to change your plans at each step to meet local concerns. Once you meet and brief your POC, you can begin coordinating.

Coordinating. The closer the time to begin data collecting, the more frequently you will speak to the POC. Expect to speak daily when you get within 30 days of data collection. In some instances, you may have to make a trip to the installation for a final coordination meeting. Be prepared to be very flexible with regard to the installations' internal schedule.

Operating. Most of the lessons learned in this category have to do with hands-on testing.

1. Many instances of equipment variation can be (and were) anticipated. Test developers and site coordinators must find out what major pieces of equipment are not likely to be available at the selected sites in advance of actual testing if high quality tracked tests are to be prepared.

2. Printed scoresheets must be proofed carefully to ensure that for every step which should be scored, a score can be recorded.

3. Scorers must be thoroughly trained, not only on how to set up and administer the tests, but also on how to record data on the scoresheets. They must be given practice in using
the scoresheets (not just talked through it) before testing, and monitored closely during testing, especially with the first few soldiers tested. Continual monitoring must also occur throughout the testing.

4. Scorers and hands-on managers must document meticulously who was tested on what, and also who wasn't tested on what, and why.

5. Experienced hands-on managers are often able to implement procedures to deal with equipment malfunctions or variations, but these too must be documented.

6. Completed scoresheets must be checked as soon as possible after testing so that careless or incorrect scoring can be detected, and the errant scorer can be retrained.

Collecting. Many of the problems that we encountered in Linking could have been avoided if all of the data were checked carefully at the site before sending them off to scanning company. We also found that where we could use a single header sheet for a group of instruments that can be scanned together, then there are fewer opportunities for discrepancies.

Processing. Never try to do too many all at once! Deal with one instrument at a time before merging. Frequently, problems will get complicated after merging.

Imputing. The decision rules and imputation procedures used with the CV data were successful in allowing us to develop performance scores for a very high proportion of the soldiers tested. Based on the available evidence, we have no reason to believe that any significant distortions were introduced while achieving this goal. Relatively few values were imputed at all. Where imputation was necessary, it was done with great care.

The apparent ease of imputation procedures should not, however, lead us to relax our data collection procedures in the future. Lessons learned from investigation of the reasons for missing data will be used to modify data collection procedures for the Project A longitudinal validation so as to further reduce the amount of missing data.

Figure 1

CONCURRENT VALIDATION TESTING BLOCKS (FOUR HOURS EACH)

BATCH A MOS		BATCH Z MOS	
Block 1	Predictor Tests	Block 1	Predictor Tests
Block 2	School Knowledge Tests MOS-Specific Job Knowledge Tests	Block 2	School Knowledge Tests Army-Wide Ratings
Block 3	MOS Specific Hands-on Tests		
Block 4	MOS-Specific Ratings Army-Wide Ratings		

Figure 2

CONCURRENT VALIDATION SAMPLE SOLDIERS BY MOS BY LOCATION

BATCH A **BATCH Z**

MOS / Location	11B	13B	19E	31C	63B	64C	71L	91A	95B		12B	16S	27E	51B	54E	55B	67N	76W	76Y	94B	Total	% Total
Fort Benning	45	23	41	7	13	30	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	66	46	0	0	37	25	41	10	72		82	75	13	19	72	20	7	42	30	62	730	7.74
Fort Campbell	90	20	0	20	66	45	54	44	43		90	23	10	0	32	18	42	51	61	46	757	8.03
Fort Carson	66	50	77	30	40	53	30	33	46		40	57	13	0	25	7	0	23	40	47	689	7.31
Fort Hood	28	56	0	30	40	20	30	50	60		51	60	4	12	62	36	44	72	41	57	767	8.13
Fort Knox	20	32	111	10	30	40	22	45	31		43	10	0	0	0	12	0	10	20	34	524	5.56
Fort Lewis	75	40	13	11	43	40	23	27	56		27	25	1	11	51	31	20	40	41	30	631	6.69
Fort Ord	30	0	0	14	30	42	31	43	51		51	7	0	1	4	7	15	23	40	20	426	4.51
Fort Polk	73	47	10	20	47	47	10	40	44		60	45	0	0	10	7	23	20	51	35	640	6.87
Fort Riley	30	43	55	27	20	45	35	30	40		31	20	0	0	25	52	0	20	30	45	570	6.14
Fort Sill	0	100	0	20	43	51	44	0	20		42	11	0	0	0	0	15	7	35	32	437	4.63
Fort Stewart	44	40	30	17	20	51	31	45	45		30	30	0	0	17	20	20	44	34	35	617	6.54
USAREUR	132	122	120	130	122	121	114	110	110		120	70	61	41	90	54	63	105	134	113	1063	20.0
Total	702	667	563	366	637	606	514	501	602		704	470	147	100	434	201	270	490	630	612	9430	
% Total	7.44	7.07	5.33	3.80	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	6.60	6.40		

Figure 3
REASONS FOR MISSING DATA

KNOWLEDGE TEST

- Soldiers Not Available for Part or All of Scheduled Time
- Soldiers Exceptionally Slow in Taking Test
- Soldiers Not Following Instructions

RATING DATA

- No Suitable Raters Available
- Soldier Does Not Perform Some Kinds of Tasks
- Rater Not following Instructions

HANDS-ON DATA

- Anticipated Variation in Equipment
- Unanticipated Variation in Equipment
- Soldiers Not Available for Part or All of Scheduled Time
- Equipment Breakdown or Nonavailability
- Conditions Preventing Testing of Some Soldiers on Some Tasks
- Scorer or Scoresheet Errors

TABLE 1

Concurrent Validation Data Base

	TOTAL RECORDS	TOTAL BYTE OF INFO[2]	TOTAL DATA POINTS[2]
Predictor Data: Paper-&-pencil	88,669	7,241K	6,321K
Predictor Data Computer	687,830	51,336K	11,832K
Criterion Data: Hands-on	77,921	3,326K	2,015K
Criterion Data: All other, including written tests	107,561	12,994K	9,283K
TOTAL	**961,981**	**74,900K**	**29,451K**

[2]Included all identifying information from each instrument

Table 2

NUMBER OF CASES WITH COMPLETE DATA FOR EACH COMBINATION OF CRITERION INSTRUMENTS

Batch A MOS

		Complete SK & JK	Comp SK Miss JK	Miss SK Comp JK	Missing SK & JK	TOTAL
Complete HO & Complete RA	N	772	189	122	58	1141
	%	14.65	3.59	2.32	1.10	21.66
Complete HO & Missing RA	N	526	130	72	29	757
	%	9.98	2.47	1.37	0.55	14.37
Missing HO & Complete RA	N	1436	364	215	125	2140
	%	27.26	6.91	4.08	2.37	40.62
Missing HO & Missing RA	N	784	241	125	80	1230
	%	14.88	4.57	2.37	1.52	23.35
TOTAL	N	3518	924	534	292	5268
	%	66.78	17.54	10.14	5.54	100.00

Table 3

NUMBER OF CASES MISSING EACH INSTRUMENT

	11B	13B	19E	31C	63B	64C	71L	91A	95B
Total N	702	667	503	366	637	686	514	501	692
Final Counts After Stage I Screening and Imputation									
Missing Hands-On	20	55	29	25	68	46	20	5	27
Missing Job Know	24	29	44	40	41	18	13	18	29
Missing Sch Know	18	28	18	17	25	17	21	22	18
Missing AW BARS	7	2	1	8	12	8	11	3	0
Missing MOS BARS	9	12	3	9	18	13	23	8	0
Missing Comb Pred	7	2	1	8	12	8	11	3	0
Missing A1: Awards	14	24	13	13	11	12	14	11	4
Missing A1: Phys Red	63	93	53	30	80	81	60	59	57
Missing A4: Arts. 15	23	28	16	14	11	14	15	14	4
Missing A5: Prom Rt	109	143	83	62	97	86	79	61	84
Total Complete	512	406	335	241	411	486	355	374	513
% Complete	72.9	60.9	66.6	65.9	64.5	70.9	69.1	74.7	74.1
Final Counts After Stage II Imputation									
Total Complete	693	656	490	356	615	675	506	492	686
% Complete	98.7	98.4	97.4	97.3	96.6	98.4	98.4	98.2	99.1

CHARACTERISTICS OF BIODATA ITEMS AND THEIR RELATIONSHIP TO VALIDITY

Bruce N. Barge

Personnel Decisions Research Institute

Presented on Symposium,

"The Use of Biodata in the 1980s and Beyond"

At the Annual Convention of the
American Psychological Association
New York

August 1987

Abstract

To investigate factors related to the predictive ability of biodata, a set of biodata items was evaluated in terms of three hypothesized characteristics: heterogeneity, behavioral discreteness, and behavioral consistency. Evaluations for each item were correlated with the validity of the item, which had been previously obtained in a large criterion-related study predicting both job performance and training performance. Results suggest that the item characteristics can be rated reliably and that the ratings are significantly related to validities. Implications of the study are discussed for both conceptual understanding of biodata and increased predictability for biodata measures.

Characteristics of Biodata Items and Their Relationship to Validity

Biographical data (biodata) has been widely used for many years and has yielded impressive validities in predicting performance in applied settings. Ghiselli (1955) found biodata to be among the best predictors available in his extensive review of cognitive and non-cognitive predictors, and more recent reviews by Owens (1976), Reilly and Chao (1982), and Barge and Hough (1984) have also supported the outstanding predictive power of biodata. Despite its notable validity, however, the mechanisms through which biodata attains prediction are still poorly understood. The constructs that organize and define the domain have not been widely accepted, and characteristics of a biodata instrument that produce its validity have not been explicitly identified.

Conceptually-oriented research on biodata has increased in the last ten years, led by Owens and his colleagues who developed and validated a conceptual model and associated biodata subgroups (Owens, 1976; Owens & Schoenfeldt, 1979; Davis, 1984). Other research has addressed stability and generalizability of biodata factors (Eberhardt & Muchinsky, 1982a, Lautenschlager & Shaffer, 1987), relationships between biodata and vocational interests (Eberhardt & Muchinsky, 1982b, 1984), and linking of biodata to job analysis components (Pannone, 1984). Rational keying of biodata scales has also been researched (Matteson, 1978; Mitchell & Klimoski, 1982).

While this increased conceptual orientation has been valuable, other aspects of biodata have received relatively little attention. In particular, little is known about the characteristics of biodata measures that contribute to their predictive validity. Many developers of biodata instruments have their own theories of what produces a valid inventory, but these theories are typically informal and have usually not been tested empirically. The purpose of this research is to delineate three hypotheses pertaining to biodata item validity, and to test these hypotheses formally.

The first hypothesis relates to the heterogeneity of biodata items, the tendency for biodata items to incorporate several "pure" behavioral tendencies or traits in a single complex behavior. For example, a biodata item pertaining to performance in school may simultaneously tap intelligence, academic interest, dependability, and several other characteristics. This heterogeneity may contribute to the item's validity because most performance criteria are also heterogeneous and require the application of several characteristics simultaneously. Thus, biodata items that are maximally heterogeneous may produce optimal validity.

A second hypothesis pertains to the behavioral discreteness of biodata items. Behavioral discreteness refers to the tendency for biodata items to address a single, perhaps verifiable,

behavior rather than a more abstract or summary characteristic. For example, a biodata item may ask about the number of jobs a respondent has held rather than asking for a self-perception of stability. Since the response to a more discrete item involves less respondent evaluation and is also potentially verifiable, the information provided may be more accurate and may produce better prediction.

The final hypothesis to be tested involves behavioral consistency, or the extent to which the behavior addressed in the biodata item parallels behavior involved in the criterion. This hypothesis is related to the "sign vs. sample" distinction (Wernimont & Campbell, 1968), which suggests that predictors can function as a "sample" that very closely parallels criterion behaviors or as a "sign" that may be very different in content from criterion behavior. An assumed advantage of biodata items is that the behaviors they address are more similar to the criterion than may be the case for "sign" measures such as personality or vocational interests. This similarity results in less of an inferential leap from predictor to criterion and may therefore improve validity. Thus, biodata items developed to closely parallel criterion behaviors may produce the highest validity.

Each of the hypotheses described above has been mentioned previously as a potential reason for the high validity of biodata measures. Each hypothesized characteristic also distinguishes

biodata from other domains of assessment, since heterogeneity, behavioral discreteness, and behavioral consistency are more often characteristics of biodata items than of items from related domains, e.g. personality or vocational interests. The characteristics may describe quite well the biodata domain as a whole, yet individual biodata items may also differ a great deal in their standing on each of the characteristics. The objective of this research was to determine whether item differences on the characteristics can be evaluated reliably and whether evaluations of these characteristics are related to the criterion-related validity of the items.

Method

Biodata items

103 items from Owens' Biographical Questionnaire (BQ; Psychological Corporation, Copyright, 1987) were evaluated in this investigation. The BQ is among the most heavily researched biodata inventories in existence and is composed of the items that were found to measure the biodata domain best in a series of iterative analyses (cf. Owens & Schoenfeldt, 1979). Items pertain primarily to experiences occurring early in life and during and shortly after high school, which was appropriate for the sample from which the validity data were available. Items were also quite diverse in content and addressed virtually all significant areas of life experience. A multiple choice, continuous response

format was used with the items, (e.g. "How active have you been in athletics? Extremely Active, Very Active, etc.").

Raters

Two groups of raters were included in the investigation, referred to as experts and students. The expert group includes 40 of the country's most highly recognized and experienced biodata researchers, each of whom has published in the area and many of whom have developed and validated biodata inventories personally. Although all members of this group could be considered expert, they differ widely in perspective. Some are employed in industry, some at universities, some at research firms, and a few are retired. In addition, their ideas about biodata are often strikingly different (based on their published work), ranging from a highly conceptual orientation to a strict empirical prediction stance.

The student group includes 17 graduate students in industrial/organizational psychology at the University of Minnesota. At the time of the investigation, students had completed between 6 months and 3 years, 6 months of graduate education, but none had extensive experience or knowledge regarding biodata in general or the development of a biodata inventory. The students differed somewhat in amount of experience, both in validation research and in research in general.

Rating of Characteristics

Rating packages were developed for each rater, including an explanation of each hypothesized item characteristic, a set of instructions and example item ratings, and a rating booklet. The rating booklet collected three ratings (one for each hypothesized characteristic) for each of the 103 items. Raters were asked to read through the rating materials, make their ratings, and return their completed booklets. They were also encouraged to bring up any questions or comments concerning the rating task.

Validity Data

Item validity data had been obtained for each of the 103 biodata items as part of Project A, which is a very large criterion-related validation effort intended to improve the selection and classification of Army enlisted personnel (Campbell, 1987; Eaton, 1984). The validity sample included junior Army enlisted personnel working in each of four jobs (radio teletype, armor crewman, vehicle operator, and administrative specialist), and validity information was available for both training and job performance criteria. Sample sizes range from 700 to 2200 for each job in the training performance data set and from 140 to 268 per job in the job performance data set.

Within each of these data sets, item validities are reported separately for a number of criterion constructs or measures developed as part of Project A. For example, the end-of-training

criteria for administrative specialists includes: 1) Typing learning rate, 2) Final typing speed, 3) Typing tasks, times tested, and 4) Nontyping tasks, times tested. Job performance criteria include five performance constructs for all jobs: Core Technical Proficiency, General Task Proficiency, Effort/Leadership/Self-development, Personal Discipline, and Physical Fitness and Personal Appearance.

Procedure

Ratings of the biodata items were analyzed separately for expert and student groups. Means, standard deviations, and frequencies were computed for each item for each characteristic. Inter-rater reliabilities were also calculated for each of the rated characteristics, separately for the individual rater and for the group composite. Finally, correlations were computed between item means for each of the characteristics, to examine both the relationships between the characteristics and the relationship between expert and student ratings.

In the validity data, the range, mean, and standard deviation of the item validities within each data set (i.e. for each criterion measure for each job) were calculated. These analyses provide information concerning the general level of validity attained by the items, as well as the variance of the validities.

Correlations were computed between the mean ratings for each item for each characteristic and the validities of the items for

each criterion/criterion construct. These correlations indicate the degree to which an item's standing on a characteristic is related to its validity. All correlational analyses were conducted separately for expert and student raters; results for each group were then compared.

Results

Item Ratings

Completed rating booklets were received from 22 members or 55% of the expert group and 12 members or 71% of the student group. Comments from the raters indicated the ratings were sometimes quite difficult to make, although difficulty apparently varied across each item-characteristic judgment. Item means were slightly lower for ratings of behavioral consistency than for heterogeneity and behavioral discreteness, and rating standard deviations were around 1.0 for each of the characteristics. Overall, the ratings ranged from 1.33 to 4.83 (on a 5 point scale), and most means were between 2 and 4. The frequency information suggested that most ratings were fairly normally distributed about the mean.

Ratings of each of the characteristics were correlated somewhat with each other. Heterogeneity correlated -.27 with Behavioral Discreteness and -.35 with Behavioral Consistency. Discreteness correlated zero with Consistency. The ratings of experts correlated highly with those of students: .78 for

heterogeneity, .89 for behavioral consistency, and .97 for behavioral discreteness.

Inter-rater reliabilities for the ratings are shown in Table 1. In general, the reliabilities are quite good and suggest the ratings are sufficiently consistent to justify relating them to the item validities. Ratings appear to be most reliable for the behavioral discreteness and behavioral consistency characteristics and are less reliable for heterogeneity. Expert and student raters attained approximately equal levels of reliability.

Item Validities

The overall absolute level of the item validities ranged from a low of zero to a high of .43. Validities were higher and had more variance in the job performance data sets than in the training performance data. Mean item validities were approximately .08 with a standard deviation of about .06 in the job performance data sets and were around .04 with a standard deviation of about .03 in the training performance data sets.

Correlations between ratings and validities

Correlational results for the job performance criteria are summarized in Tables 2 through 5. These tables report correlations between item mean ratings for each characteristic and item validities for each of five performance criteria. Results are therefore an index of the relationship between ratings of an item's characteristics and the item's validity in predicting various dimensions of job performance.

Tables 2 (expert results) and 3 (student results) show the correlations obtained when computed in the sample as a whole and when computed separately within each job and then averaged. Tables 4 and 5 present similar results, reported separately for the administrative specialist job only and the armor crewman job only. The correlations obtained for the administrative specialist job are the highest of all four jobs, while results for the armor crewman are the lowest and least consistent. Despite this across-job variability in the level of correlation, the pattern of correlation for each of the jobs is quite consistent with the overall across-job average.

Several findings are notable in the correlational results. First, ratings of heterogeneity are negatively correlated with the validity of the items; that is, items that were rated low on heterogeneity are more valid than items judged to be high in heterogeneity. Ratings of behavioral discreteness and behavioral consistency are positively correlated with item validities. Thus, items rated as discrete and consistent with criterion behavior tend to produce higher validities than items of a more evaluative or summary nature or items that function as a "sign" of criterion behavior rather than a "sample".

The magnitude of the relationships between ratings and validities is strongest for the criteria of Core Technical Proficiency and General Task Proficiency. These performance dimensions are known informally as the "can-do" criteria, as opposed to "will-do" criteria such as Effort/Leadership, Personal

Discipline, or Physical Fitness. Relationships are also much stronger when computed within the administrative specialist job alone. Finally, comparison of results shows that both the pattern and level of correlation is highly consistent for both expert and student raters.

Findings from correlational analyses with the training performance validities present a similar picture, as shown in Tables 6 and 7. Results are averaged across the training criteria within each job, since the criteria differed by job and it was therefore impossible to compare results across jobs. Averaging across training criteria within a job also appears reasonable since the criteria are similar conceptually (e.g. typing learning rate and final typing speed).

As with the job performance criteria, correlations are negative between training validities and ratings of heterogeneity and are positive between validities and ratings of behavioral discreteness and behavioral consistency. The strongest relationships are for the vehicle operator and administrative specialist jobs, a result that is also found in the job performance results. Finally, the pattern and level of correlation obtained is again highly similar for expert and student raters, although the expert ratings attained slightly higher relationships with the training validities.

Discussion

Although preliminary, investigation results suggest that each of the three hypothesized characteristics (heterogeneity, behavioral discreteness, and behavioral consistency) is an important, stable, descriptor of biodata items and their ability to predict criteria. Each characteristic was rated reliably by both expert and student raters, and each characteristic correlated significantly with item validities for both job performance and training performance criteria. Behavioral consistency appears to be the item characteristic of most value in predicting an item's validity, especially for job performance criteria, but both heterogeneity and behavioral discreteness also attained respectable correlations with the item validities.

The direction of the relationship between characteristics and validities is as hypothesized for behavioral discreteness and behavioral consistency, suggesting that items that are both behaviorally discrete and consistent with criterion behavior are likely to yield the best validities. For heterogeneity, the relationship is opposite to that hypothesized, suggesting that items that are less heterogeneous are more likely to produce validity. This finding is interesting since heterogeneity was examined at the item level rather than at the scale or inventory level as is more traditional in research. Thus, it may be that heterogeneity is still desirable in a biodata instrument, but that

such heterogeneity is best attained by combining items that are themselves somewhat homogeneous. The ratings of heterogeneity were also noticeably less reliable than for the other characteristics, and while this should not affect the direction of its relationship with validities, the heterogeneity characteristic may be the most difficult of the characteristics to interpret.

It is interesting that the relationship between characteristics and validities is notably stronger in the administrative specialist job than it is across jobs. This is also true for the job performance criteria of Core Technical Proficiency and General Task Proficiency. A possible explanation for the administrative specialist finding is that the validities for this job were obtained in a sample that was the largest of the jobs included (N = 268 for job performance criteria and N = 2260 for training criteria). The sample sizes in the other jobs are all at least reasonably large, however, so it appears other factors are also involved. For the job performance criteria, the two dimensions that are well predicted are both referred to as "can-do" criteria, while the other dimensions are "will-do" criteria. Again, however, it is not clear why this finding should be obtained. Future research to extend the findings of this investigation should address the stability of characteristic-validity relationships across both jobs and criteria.

Because this research is the first investigation attempted of

biodata item characteristics and their relationship to validity, results must be viewed with caution. Only one set of biodata items (aimed primarily at young adults) was included and the criterion-related validities were gathered in the military rather than in an industrial organization. Nevertheless, several strengths of the investigation suggest the findings may be relatively stable in future research.

First, the item set employed is highly diverse, which should contribute to an effective test of the rated characteristics. Second, the level of validities attained by the items is respectable, and even more important, the validities have considerable variance. The validities were obtained in large "real-world" samples, using criteria that had been carefully developed. Finally, the results obtained are highly consistent for two independent groups of raters and are also consistent for two independently gathered types of criteria that are conceptually and methodologically distinct.

Results from this research can be examined from both theoretical and applied perspectives. From the theoretical perspective, this investigation suggests conceptual reasons that may underlie the predictive ability of biographical measures. Item characteristics examined in the research are more often characteristics of biodata items than of items from related domains; they may therefore be in part responsible for the

superiority of biodata prediction to that from other domains. The investigation's findings also complement more content-related biodata research, such as that addressing biodata factors like academic achievement or early home environment. A taxonomy combining both the characteristics examined in this research and content-related aspects of biodata measures may be of great value in improving conceptual understanding of the biodata domain.

From the applied perspective, findings of the investigation address the optimal construction procedures for a biodata instrument. Items developed with attention to the characteristics studied may produce higher validities, an outcome of obvious applied value. This value may be increased further through combination of content considerations with the characteristics examined here.

A conference in 1965 of the nation's leading biodata researchers concluded:

> Aside from theoretical academic interest, there were no very persuasive reasons for tackling the (biodata conceptual) problem until a 'prediction plateau' developed. It seems apparent now that increased efficiency will occur only when we learn more about the causal relationships underlying predictive items (Henry, 1966, p. 248).

Future research to replicate and extend both theoretical and practical aspects of this investigation will hopefully be of value in contributing to this continuing goal.

Table 1

Inter-rater Reliabilities for Judgments of Biodata Item Characteristics

	Heterogeneity		Behavioral Discreteness		Behavioral Consistency	
	Experts	Students	Experts	Students	Experts	Students
Within Group	.79	.72	.95	.95	.92	.83
Individual Rater	.15	.17	.49	.61	.36	.27

Expert Sample N = 22

Student Sample N = 12

Table 2

Correlation Between Expert Ratings and Validities: Job Performance

	Computed for Total Sample (N = 746)				
	Core Technical Proficiency	General Task Proficiency	Effort, Leadership, and Self-Development	Personal Discipline	Physical Fitness and Military Appearance
Heterogeneity	-.33**	-.44**	-.06	-.01	-.06
Behavioral Discreteness	.19*	.19*	-.07	-.09	-.04
Behavioral Consistency	.35**	.43**	.33**	-.10	.36**

	Computed for Each Job and Averaged (N = 140 - 268 per job)				
	Core Technical Proficiency	General Task Proficiency	Effort, Leadership, and Self-Development	Personal Discipline	Physical Fitness and Military Appearance
Heterogeneity	-.14	-.17	-.04	.02	-.08
Behavioral Discreteness	.16	.16	-.03	.03	-.05
Behavioral Consistency	.19*	.26**	.20*	.00	.32**

** $p < .01$

* $p < .05$

Table 3

Correlation Between Student Ratings and Validities: Job Performance

	Computed for Total Sample (N = 746)				
	Core Technical Proficiency	General Task Proficiency	Effort, Leadership, and Self-Development	Personal Discipline	Physical Fitness and Military Appearance
Heterogeneity	-.30**	-.41**	-.08	.03	-.12
Behavioral Discreteness	.16	.18	-.09	-.07	-.04
Behavioral Consistency	.34**	.43**	.15	-.20*	.25**

	Computed for Each Job and Averaged (N = 140 - 268 per job)				
	Core Technical Proficiency	General Task Proficiency	Effort, Leadership, and Self-Development	Personal Discipline	Physical Fitness and Military Appearance
Heterogeneity	-.14	-.19*	-.02	.04	-.12
Behavioral Discreteness	.15	.18	-.04	.04	-.01
Behavioral Consistency	.19*	.28**	.12	-.06	.24*

** $p < .01$

* $p < .05$

Table 4

Correlation Between Expert Raters and Validities: Job Performance

For Administrative Specialist Job Only (N = 268)

	Core Technical Proficiency	General Task Proficiency	Effort, Leadership, and Self-Development	Personal Discipline	Physical Fitness and Military Appearance
Heterogeneity	-.33**	-.35**	-.19*	-.21*	-.09
Behavioral Discreteness	.33**	.26**	.15	.34**	.00
Behavioral Consistency	.22*	.54**	.38**	.10	.33**

For Armor Crewman Job Only (N = 176)

	Core Technical Proficiency	General Task Proficiency	Effort, Leadership, and Self-Development	Personal Discipline	Physical Fitness and Military Appearance
Heterogeneity	.13	-.26**	.12	.10	-.12
Behavioral Discreteness	-.02	.29**	-.20*	-.17	-.02
Behavioral Consistency	.01	.16	.13	.00	.39**

** $p < .01$

* $p < .05$

Table 5

Correlation Between Student Raters and Validities: Job Performance

	Core Technical Proficiency	General Task Proficiency	Effort, Leadership, and Self-Development	Personal Discipline	Physical Fitness and Military Appearance
	For Administrative Specialist Job Only (N = 268)				
Heterogeneity	-.26**	-.41**	-.27**	-.21*	-.12
Behavioral Discreteness	.34**	.26**	.17	.35**	-.02
Behavioral Consistency	.23*	.60**	.39**	.11	.25**
	For Armor Crewman Job Only (N = 176)				
Heterogeneity	.11	-.22*	.21*	.14	-.21*
Behavioral Discreteness	-.02	.31**	-.21*	-.14	-.04
Behavioral Consistency	.03	.13	-.05	-.13	.30**

** p < .01

* p < .05

Table 6

Average Correlation Across Criterion Measures Between Expert Ratings and Validities: Training Performance

	Radio Teletype N = 726	Armor Crewman N = 1642	Vehicle Operator N = 1076	Administrative Specialist N = 2260
Heterogeneity	-.22*	-.11	-.26**	-.29**
Behavioral Discreteness	.09	.13	.31**	.35**
Behavioral Consistency	.12	.12	.28**	.31**

Table 7

Average Correlation Across Criterion Measures Between Student Ratings and Validities: Training Performance

	Radio Teletype N = 726	Armor Crewman N = 1642	Vehicle Operator N = 1076	Administrative Specialist N = 2260
Heterogeneity	-.13	-.10	-.17	-.26**
Behavioral Discreteness	.08	.13	.21*	.36**
Behavioral Consistency	.10	.12	.29**	.27**

** $p < .01$

* $p < .05$

References

Barge, B. N., & Hough, L. M. (1984). Utility of biographical data: A review and integration of the literature. Minneapolis, MN: Personnel Decisions Research Institute.

Campbell, J. P. (Ed.). (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1986 fiscal year. (HumRRO IR-PRD-87-10) (ARI Technical Report in preparation).

Davis, K. R. (1984). A longitudinal analysis of biographical subgroups using Owens' developmental-integrative model. Personnel Psychology, 37, 1-14.

Eaton, N. K. (1984, May). The U. S. Army research project to improve selection and classification decisions. Paper presented at the National Security Industrial Association Conference on Personnel and Training Factors in Systems Effectiveness, Springfield, VA.

Eberhardt, B. J., & Muchinsky, P. M. (1982a). An empirical investigation of the factor stability of Owens' Biographical Questionnaire. Journal of Applied Psychology, 67, 138-145.

Eberhardt, B. J., & Muchinsky, P. M. (1982b). Biodata determinants of vocational typology: An integration of two paradigms. Journal of Applied Psychology, 67, 714-727.

Eberhardt, B. J., & Muchinsky, P. M. (1984). Structural validation of Holland's hexagonal model: Vocational classification through the use of biodata. Journal of Applied Psychology, 69, 174-181.

Ghiselli, E. E. (1955). The measurement of occupational aptitude. Berkeley, CA: University of California Press.

Henry, E. R. (Chrmn.) (1966). Research conference on the use of autobiographical data as psychological predictors. Greensboro, NC: The Creativity Research Institute, The Richardson Foundation.

Lautenschlager, G. J., & Shaffer, G. S. (1987). Reexamining the component stability of Owens' Biographical Questionnaire. Journal of Applied Psychology, 72, 149-152.

Matteson, M. T. (1978). An alternative approach to using biographical data for predicting job success. Journal of Occupational Psychology, 51, 155-162.

Mitchell, T. W., & Klimoski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. Journal of Applied Psychology, 67, 411-418.

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally.

Owens, W. A., & Schoenfeldt, L. F. (1979). Towards a classification of persons. Journal of Applied Psychology Monograph, 64, 569-607.

Pannone, R. D. (1984). Predicting test performance: A content valid approach to screening applicants. Personnel Psychology, 37, 507-514.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62.

A TASK-BASED APPROACH FOR IDENTIFYING JUNIOR NONCOMMISSIONED OFFICERS' KEY RESPONSIBILITIES

Ilene F. Gast
U.S. Army Research Institute

Charlotte H. Campbell
Human Resources Research Organization

Alma G. Steinberg
U.S. Army Research Institute

Daniel A. McGarvey
American Institutes for Research

Presented on Symposium,

"Junior Noncommissioned Officer Job Requirements:
Where Does Leadership Fit In?"

At the Annual Convention of the
American Psychological Association
New York

August 1987

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

This research was funded by the U.S. Army Research institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

A Task-Based Approach for Identifying Junior NCOs' Key Responsibilities

Introduction

Project A, Improving the Selection and Classification of Army Enlisted Personnel, has begun to address the prediction of long range criteria. One of the challenges faced by project personnel is the development of appropriate measures of second tour job knowledge and hands-on performance. As soldiers enter their second tour of duty, their jobs change dramatically. During their second tour, soldiers begin assuming supervisory responsibilities, while also retaining their technical duties. As Dr. Rumsey mentioned, unlike technical tasks, many of these new supervisory activities cannot be translated into discrete, proceduralizable tasks and therefore are not amenable to the same kinds of job analytic procedures. For similar reasons, the measurement strategies developed for the technical first-tour tasks, hands-on and job knowledge tests covering specific tasks, would capture second tour performance insufficiently. For these reasons, we designed a job analytic approach incorporating multiple data sources to determine the appropriate mix of supervisory and technical skills, ensure adequate coverage of both domains, and provide insight into suitable measurement procedures.

Today, I will describe how we integrated data from three sources to develop a comprehensive performance domain for each of nine Army jobs. I will then describe some general differences we found in the composition of first and second tour jobs. In addition, I will discuss any specific

differences among the selected Army jobs in the importance of supervisory activities. Finally, I will address practical implications for the development of "task-based" second tour performance measures.

Method

Developing the Task Domains

Our first step was assembling a comprehensive list of job-specific tasks within each of nine Army occupations. Included were: infantryman (11B), cannon crewmember (13B), tank crewmember (19E), single-channel radio operator (31C), light wheel vehicle mechanic (63B), motor transport operator (64C/88M), administrative specialist (71L), medical specialist (91A/B), and military police (95B). Three separate job analytic procedures were employed to generate tasks to be included in this list. Brief descriptions of each are provided in the following section.

Technical Components of the Junior NCO Job

I'll begin by discussing the process we used to describe the technical portion of junior non-commissioned officers' (NCOs') jobs. For each of the nine jobs being studied, definition of the junior NCO job domain began with the Soldier's Manuals for the job. Soldier's Manuals are prepared by Army agencies for every job and for every skill level within the job. They not only list the tasks required, but also the conditions under which they are performed, the steps required for performance, and the performance standards. The Army also expects soldiers to be proficient on the tasks in the Soldiers' Manual of Common Tasks, which likewise includes tasks, conditions, steps, and standards for basic soldiering tasks at each skill level (tasks such as map

reading, basic first aid, and operation of individual weapons). The junior NCO, who is a Skill Level 2 soldier, is held responsible for all Skill Level 1 and 2 tasks in both the job-specific and common task manuals.

We also used data from the Army Occupational Survey Programs (AOSPs) in defining the job domains. These surveys, which list hundreds of task statements for each job, are administered periodically to represent samples of soldiers at every skill level of each job; analyses of the data include the percent of soldiers at each skill level who report that they perform the tasks. The list was screened to eliminate statements not performed by Skill Level 2 soldiers, and the surviving AOSP statements were then mapped on to the tasks list defined by reference to the Soldier's Manuals. Any AOSP statements that were not thus subsumed under Soldier's Manual tasks were added to the domains. In so doing, we often found that higher skill level tasks were performed by significant numbers of Skill Level 2 soldiers, and were therefore considered to be a part of the job domain in hand.

With this domain list in hand, we visited the Army agencies responsible for training and doctrine in each job, and requested their review of the list. We asked for their input concerning the completeness and accuracy of the list, and also found out from them whether any of the tasks were likely to be eliminated soon because of equipment or doctrine changes, or whether other tasks should be added for similar reasons. After they had completed their review and given their concurrence on the doctrinal accuracy of the domain, we considered domain definition of the technical tasks complete. (The process parallels that used in defining the Skill Level 1 domains, for

which only Skill Level 1 Soldier's Manuals were used in the initial step. Details may be found in Campbell, Campbell, Rumsey & Edwards, 1985).

Supervisory Components of The Junior NCO Job

In developing the supervisory portion of the job domains, we took advantage of existing research on Army supervision and leadership. Two separate approaches were incorporated to form the supervisory component of the job domains. The first approach yielded the Supervisory Responsibility Questionnaire (SRQ). It was based on critical incidents describing working relationships between first term soldiers and their NCO supervisors. The second approach was an adaptation of the Leader Requirements Survey (LRS), an extensive interview-based task list. By combining these two instruments we were able to take advantage of two job analytic techniques and two different slices of the supervisory domain. Next, I will discuss the development of the SRQ and LRS and describe how they were combined to define the second tour supervisory domain.

The Supervisory Responsibility Questionnaire (SRQ): A Behavioral Example Based Task list. The SRQ was the byproduct of previous Project A research which examined possible moderating effects of supervision on the relationship between soldiers' pre-enlistment attributes (such as aptitude and temperament) and their job performance (White, Gast & Rumsey, 1985; Hough, Gast & White, 1986). As part of this research, critical incidents had been collected in order to determine what supervisory behaviors made a difference to the performance of first tour soldiers.

The incidents were written by 80 NCO subject matter experts (SMEs) from five of the Project A target Jobs (11B, 19E, 31C, 63B, and 91A). The SMEs were asked to recall from their experiences as first tour soldiers examples of how their supervisors had been particularly effective or particularly ineffective. In all the SMEs generated over 400 behavioral examples. Next, a retranslation was conducted in which a second group of 31 SMEs, who were familiar with Army leadership requirements, were asked to classify the examples into Yukl's 13-dimension taxonomy of supervisory behavior (Yukl, 1984). At the same time, these examples were classified by two ARI staff psychologists. As a result of retranslation, 9 of the 13 Yukl dimensions remained. (See Table 1)

The SRQ was constructed from a subset of these incidents and their respective categories. First, all incidents that were not categorized into a single dimension by both SME's and psychologists were eliminated from consideration as were multiple incidents referring to a single task or behavior. The incident list was further reduced by excluding incidents not describing a specific task (e.g., "The soldier fell asleep while on guard duty. [The supervisor] walked up to the sleeping soldier and scared him.") In the end, a total of 34 behavioral statements were written to represent 8 of Yukl's original categories. No statements were written for one category, Act as Role Model, because the incidents grouped under that category were not rich enough to extract critical supervisory tasks.

One interesting facet of the SRQ is the use of critical incidents as the basis for task statements, which is not typical. The critical incident

technique, a behavioral job analytic procedure, is typically used for developing broader behavioral dimensions. (See Pulakos, Hanson, Borman, Hallam, Carter & Owens-Kurtz, 1987).

Because the incidents comprising the SRQ tasks were primarily concerned with relationships between supervisors and their subordinates, the SRQ had built-in limitations in its coverage of the supervisory domain. To ensure all important aspects of the supervisory domain were included, a supplemental task list was needed.

The Leader Requirements Survey (LRS) Interview-based Task list. The second approach incorporated the Leader Requirements Survey (LRS), which was originally designed to provide the Army's proponents for leadership with information about the leadership job requirements of Army commissioned and non-commissioned officers. The LRS was designed to identify the sequential and progressive nature of commissioned and noncommissioned officer leadership (second lieutenant through colonel, and sergeant through command sergeant major), and contains items which cover the leadership domain of all these organizational levels. In addition, it embodies the full range of leadership tasks across all Army branches.

This task list was constructed through an iterative interview strategy. Several hundred interviews were conducted. Typically, 6-8 SMEs were interviewed at a time and interviews lasted for approximately 90 minutes. Interviewees were asked to describe their job, focussing particularly on what they did to influence others to accomplish their mission (i.e., the Army definition of military leadership as documented in FM 22-100) and especially

those leadership tasks that differentiated their jobs from those performed by others in higher or lower ranks than themselves.

In order to ensure that the resulting task list both completely encompassed the domain of military leadership and was worded in terms commonly employed by job incumbents, each successive group of SMEs was shown the leadership tasks developed by the previous groups and asked to comment on these tasks. These iterative interviews were conducted until new groups of SMEs no longer added new tasks and were comfortable with the wording of tasks already collected.

Content validation of the task list was achieved through reviews by two separate groups of Army leadership proponents, The Center for Army Leadership (CAL) and The U. S. Army Sergeants Major Academy (USASMA). Consensus on the final list of tasks comprising the LRS was reached by a review committee consisting of representatives from CAL, USASMA and ARI. The resulting Leader Requirements Task List contains tasks in the following broad categories: Train, Teach, and Develop (146 tasks); Motivate (170 tasks); Manage (86 tasks); and Provide Direction (158 tasks) for a total of 560 tasks. Table 2 lists the number of tasks corresponding to each content area. In the present research, 25 of the tasks in the category "Provide Direction", coming under the sub-heading of "Provide Input for the Direction of the Larger Organization" were dropped because they contained tasks performed by higher-level commissioned officers only. (See Steinberg, 1986, and Steinberg, van Rijn & Hunter, 1986 for more information on the LRS). A listing of number of tasks by content area can be found in Table 2.

Combining the LRS and SRQ

The Supervisory Responsibility Questionnaire gave us information on working relationships between first-line supervisors and their subordinates, as perceived by the subordinates. The much longer Leader Requirements Survey included activities involving peers and superiors, as well as administrative duties, but was designed to cover these activities across all supervisory levels within all Army branches, from junior NCOs through full Colonels. In order to determine which of the activities on the LRS should be a part of the domains for junior NCOs, and to verify the tasks on the SRQ as appropriate tasks for the domains, both questionnaires were administered to NCOs in the nine jobs. Approximately 50 NCOs received the LRS, and 125 NCOs received the SRQ. For each questionnaire, the NCOs were asked to indicate how important each task is in performance of the E5's[2] job; on the SRQ, they were also asked to indicate how frequently each task is performed.

Analysis of the SRQ data confirmed that all the tasks were sufficiently important, across a variety of Army jobs, to be retained as part of the junior NCO domain. The LRS importance data were used to select tasks that over half of the respondents indicated were absolutely essential to the E5's job. Additional highly rated tasks were also selected from any of the 19 LRS

---

[2]E5 is the Army paygrade at which Army doctrine specifies that soldiers become noncommissioned officers and can assume supervisory responsibilities. E5 is also the first paygrade at which a soldier is classified as Skill Level 2. We were particularly concerned with the jobs of E5 soldiers because it was projected that E5 would be the most common paygrade within our second tour Project A sample, and thus was designated as the target group for whom we would develop our measures.

dimensions not already represented by at least two tasks. Ultimately, two LRS dimensions were eliminated from the domain because they failed to meet the importance criteria. By this process, 53 tasks were selected from the LRS to be considered for the job domains.

Content analysis of the two tasks lists—34 tasks from the SRQ and 53 tasks from the LRS — resulted in a single list of 46 tasks that incorporated all of the activities on both lists. Of those 46, the 34 tasks from the SRQ were included and 8 of the LRS tasks; 4 new task statements were prepared to cover two or more LRS statements each.

The 46 Task statements were further examined by reference to the categories used for the original SRQ. Eight categories evolved for the 46 tasks, shown in Table 3. These tasks, clustered as shown, were added to the Skill Level 2 job domain for each of the nine jobs.

Refining the Job Domains

After the job domains were thus defined, every domain included over 200 tasks. We wanted to select smaller samples of tasks to represent each of the domains, samples that would include the most critical tasks for the jobs and that would have a sufficient range of performance difficulty to permit some discrimination among soldiers. In order to do this, more information was needed; specifically, we needed judgments of task criticality and performance difficulty.

The Army agency responsible for each job was asked to designate 30 job experts: officers or NCOs in that military specialty who had recent field experience supervising E5s in the job. Half of the job experts rated the

tasks for a hypothetical E5 soldier who had between three and five years of service; half were given another task not directly related to the topic under discussion today.

For the importance judgments, the experts were given one of three scenarios, and asked to rate (on a 5-point scale) the importance of the task in accomplishing the unit's mission under that scenario. The three scenarios described either combat conditions (European, non-nuclear), increasing tensions (European, with a high state of training and strategic readiness, but short of actual conflict), or a garrison environment (stateside, with training as the primary activity and mission). In all, we collected 10 ratings for each paygrade/scenario combination, for a total of 30 sets of ratings per job. The importance ratings were averaged across the 10 experts in each rating condition to yield 3 importance scores. To obtain an indication of expected task performance distribution, the experts were asked to sort a "typical" group of ten hypothetical soldiers into five performance categories based on how they would expect soldiers to be able to perform on each task. Task difficulty was then computed as the mean of the distribution of the ten soldiers, averaged across the experts. Task performance variability was computed as the standard deviation of the distribution of the ten soldiers, averaged across experts. (This procedure, for both importance and difficulty ratings, was developed and used for the Skill Level 1 job analysis, and is described in more detail in Campbell et al., 1985.)

Selecting Tasks for Measurement

The last step before designing hands-on and job knowledge measures for each of the job domains is selecting a subgroup of tasks for measurement. Even as we speak, that task selection process is taking place. The Army agencies for each job have again been asked to provide six job experts with recent field experience; one Project A staff member will also serve on the task selection panel. The information to be presented for their consideration includes the tasks, clustered; the importance rating for each task for E5s; the performance difficulty and variability for each tasks for E5s; and the performance frequency for each task, drawn from the Army Occupational Survey Program analyses. The panel will eventually agree on 45 tasks to be selected for each job, 30 technical tasks and 15 supervisory tasks. To guide their selection so that every cluster is represented, targets are set proportionally for each cluster.

Analyses

The analyses addressed general differences between first and second tour jobs and differences in supervisory requirements across jobs. Two sources of data were considered (1) the second tour job analysis data described in this paper and (2) clusters derived from first tour job analyses described by Campbell et al, 1985.

Analyses were largely descriptive because with the exception of the SRQ data, domains were not directly comparable between first and second tour nor across jobs. Further, these analyses were preliminary; as we move from task selection to task measurement we will be using these data to answer questions

about the best way to capture performance on specific tasks. In order to examine general differences between first and second tour domains, we began with a general comparison of the content of these two sets of domains for each MOS. We noted trends for changes in cluster composition, the addition of new clusters, and the deletion of existing clusters from first to second tour. To examine differences among the occupations in supervisory responsibilities we assessed job-specific additions to the core SRQ clusters and importance ratings for each of the augmented clusters. Finally, we compared differences in importance ratings across occupations for the common, technical and supervisory cluster groupings. Our research questions follow.

1. How much overlap is there between first and second tour job dimensions?

2. Do jobs increase in complexity? What are the indications based on comparison of first and second tour domains?

3. What is the balance between supervisory and technical tasks in the second tour? Does this balance vary across MOS?

4. In which job(s) were supervisory activities judged to be the most important? The least important?

5. Did specific supervisory activities differ in importance across the jobs?

Results

Changes in Domains from First to Second Tour: Common and Technical Tasks

Our first step was to examine changes in clusters of "Common Tasks" which are shared to some extent across MOS, and non-shared job-specific

"technical" tasks. In terms of the dichotomy presented by Dr. Rumsey between supervisory and technical tasks, both of these would represent subsets of the "technical" category. At the time we prepared our analyses, the databases for two of the jobs were still under preparation at the time we were writing this paper. Table 4 compares the first and second tour job domains across seven jobs in terms of (1) the number of clusters included and (2) the number of tasks included. In all but two instances (infantryman and motor transport operator) the number of clusters needed to describe the domain increased from first tour to second tour jobs. Moreover, in all cases, there was an increase in the number of tasks needed to describe each domain.

Differences in technical clusters across jobs. Although first and second tour technical tasks and clusters overlap considerably, differences between the two sets of domains emerged. Specifically, within the four non-combat jobs (radio operator, motor transport operator, administrative specialists, and medical specialists), tasks in three of the shared clusters realigned themselves. By contrast, there was little change in the combat jobs, infantryman, cannon crewman or tank crewmember, with the least amount of change for the infantry.

The addition of tasks also caused several of the technical clusters to split into better differentiated groups of tasks. A more important change was the addition of an Operations (or Tactical Supervision) category in four of the seven jobs (infantry, tank crewmember, radio operator, and administrative specialist) and expansion of that category in a fifth (cannon crewmember). Further, the two jobs (radio operator and medical specialist)

not acquiring an Operations cluster gained a separate Administrative cluster. (See Table 5.)

Supervision as a Component of Second Tour Job Domains

In addition to acquiring new kinds of job specific technical responsibilities, each job acquired supervisory duties. We mentioned earlier that the 46-item SRQ was appended to each domain prior to domain review. As a result of domain review, many supervisory tasks from the AOSP and Soldier's Manuals were grouped under existing SRQ clusters. Table 5 shows the number of AOSP/SM tasks added to each SRQ cluster by occupation. The bulk of the tasks were added to three of the clusters: (1) Plan, Organize, Monitor; (2) Provide Information, and (3) Train, Develop. The remaining five SRQ clusters remained fairly stable across occupations. Within the three most augmented clusters, new SRQ tasks were not evenly distributed across the different jobs. Four of the occupations (tank crewmember, radio operator, administrative specialist and medical specialist) gained proportionally more tasks than the other occupations. Thus, domain review served to augment the SRQ, albeit unevenly.

Our next step was to assess the relative importance of the new supervisory responsibilities for soldiers in E5 paygrade across the various jobs. Past experience and preliminary analyses conducted last year suggested that not only will the overall importance of supervision vary across jobs, but specific supervisory activities vary in importance across jobs. Others postulated that E5 jobs were largely technical and that Project A need not be concerned with measuring supervisory tasks.

At the time this paper was prepared, task importance data were available for six of the nine occupations. (See Table 6.) With few exceptions, the data ran counter to our expectations. First, with the exception of the medical specialists, supervisory activities were judged to be fairly important across all jobs examined. The medical specialists' responses were consistently lower; however, if one were to add a constant of .6 to all dimensions (except discipline/punish) one would find the cluster means within the ranges produced by the other jobs.

Second, the means presented in Table 6 suggest that specific supervisory activities are not differentially important within each occupation. With one striking exception, supervisory clusters tended to have similar importance ratings within each job. Across the board, the cluster "Act as a Role Model" was first (or second in for one job) in importance. Within that category "leading by example" was the most important task.

A third unanticipated but welcome finding was the relative balance in the importance of supervisory and technical tasks across all jobs. (Again, ratings for the medical specialists were consistently lower across all facets of the domain than ratings for other occupations.)

Discussion

In many ways, second tour jobs are more complex than their first tour counterparts. As more tasks enter the job domains, the clusters became more clearly differentiated. Not only are soldiers doing more of the same types of tasks but they are also acquiring new responsibilities, particularly in the areas of task-specific supervision and administrative responsibilities.

In addition, these new supervisory responsibilities are considered to be important within each of the job domains.

The final version of the SRQ was instrumental in capturing these supervisory responsibilities. Without it we would have missed important facets of second tour jobs. The Soldier's Manuals and AOSP surveys would have done at best, an uneven job of representing the supervisory portion of the domains.

We also found that supervisory tasks were seen as fairly important within all of the jobs. While more analyses of the data collected are needed to systematically explore patterns of differences between jobs, even a cursory examination of task means by occupation reveals that many of the expected differences within and between occupations did not arise. However, what did emerge was a balance among technical and supervisory aspects of the domain. Nevertheless, based on our results, it would be difficult to conclude that supervisory activities are important within some occupation(s) and not others. Further the data provide little evidence that tasks were differentially important in specific jobs.

However, the job analyses did not provide as much information about job specific differences in supervisory activities as had been expected. Therefore, we will be totally dependent on our next phase to determine which aspects of supervision will be measured for each MOS. During task selection, SMEs will be forced to prioritize tasks with the knowledge that only the top 45 tasks will be considered for measurement purposes.

As a final note, this research is consistent with past research in

leadership and supervision. First, support is provided for Yukl's taxonomy of leadership (1984). The categories he proposed held across two different job analytic techniques and across several Army occupations. Second, the importance ratings that the SMEs gave to the "Act as Role Model" category are consonant with Bass's (1985) transformational leadership theory. According to Bass, leaders who can draw on their own informal sources of power (e.g., being a good model) to motivate others are more effective than those who rely on organizational incentives. The overwhelming agreement on the importance of leading by example typifies this kind of leadership.

Table 1

Comparison between Yukl's (1984) Categories of Supervisory Behavior and SRQ Categories

Yukl's (1984) Categories	SRQ Final Categories
1. Planning and Organizing	1. Plan, Organize, Monitor
2. Monitoring	
3. Problem Solving[a]	
4. Leading by Example	2. Act as Role Model
5. Recognizing and Rewarding	3. Recognize, Reward
6. Training and Developing	4. Train, Develop
7. Informing	5. Provide Information
8. Delegating/Participation[a]	
9. Supporting	6. Support
10. Disciplining/Punishing	7. Discipline, Punish
11. Representing[a]	
12. Promoting Teamwork[a]	
13. Clarifying Roles and Expectations	8. Clarify Roles, Provide Feedback

Notes. [a]Deleted from list as a result of retranslation (White, Gast & Rumsey, 1985)

Table 2

Leader Requirements Survey (LRS): Number of Tasks by Content Area

Content Area	Number of Tasks
TRAIN, TEACH, DEVELOP	
Train Soldiers	21
Teach Soldiers	18
Develop Leaders	21
Plan & Conduct Training	42
Train in the Field to Enter Combat	44
MOTIVATE	
Motivate Others (The What)	13
Motivate Others (The How)	42
Develop Unit Cohesion	52
Reward & Discipline Subordinates	30
Take Care of Soldiers	33
MANAGE	
Manage Resources	40
Perform/Supervise Administrative Functions	26
Coordinate with Others Outside the Unit	20
PROVIDE DIRECTION	
Supervise Others	20
Maintain Two-way Information Exchange with Subordinates	21
Maintain Two-Way Information Exchange with Superiors	17
Monitor and Evaluate Performance	38
Conduct Counseling	24
Establish Direction of the Unit/Element	13
Provide Input for the Direction of the Larger Organization	25
TOTAL	560

Table 3

Supervisory Categories and Tasks Derived From Combining the SRQ and LRS

PLAN, ORGANIZE, MONITOR
Check tools, equipment and supplies used by subordinates
Assign work tasks to subordinates
Check on subordinates during task performance
Inspect completed work
Conduct formal scheduled inspections
Monitor condition of equipment and supplies
Meet suspense dates and deadlines
Motivate subordinates in maintenance
CLARIFY ROLES, PROVIDE FEEDBACK
Provide informal feedback on task performance
Counsel subordinates - Scheduled formal counseling
Counsel subordinates - Unscheduled formal counseling
Monitor military appearance and bearing
PROVIDE INFORMATION
Brief newly assigned personnel
Answer work related questions
Conduct meetings
Brief subordinates on current of future missions/requirements

Table 3 (Continued)

Supervisory Categories and Tasks Derived From Combining the SRQ and LRS

Pass information down chain of command

Notify subordinates of changes in plans

Present formal information briefings

Provide positive input to supervisors

Provide commander with information on enemy situation

RECOGNIZE, REWARD

Provide positive verbal feedback

Reward soldiers for performance

Recommend Soldiers for promotion

Write up recommendations for awards

Integrate subordinates into the unit

TRAIN, DEVELOP

Instruct subordinates on task performance

Provide individual job training

Conduct team training

Provide remedial instruction

Encourage use of training manuals or job aids

Counsel subordinates in career planning/personal development

Table 3 (Continued)

Supervisory Categories and Tasks Derived From Combining the SRQ and LRS

Develop training plans
Provide opportunities for leadership
SUPPORT
Listen to subordinates' personal problems
Counsel subordinates on personal problems
Arrange assistance for personal problems
Assist soldiers with personal problems
DISCIPLINE, PUNISH
Issue verbal reprimands
Counsel subordinates with disciplinary problems
Arrange for extra training and disciplinary/corrective action
Recommend judicial or non-judicial action to the commander
Resolve disputes among subordinates
ACT AS A MODEL
Set the example
Remain with the assigned unit or element under adverse or wartime conditions
Share subordinates hardship

Table 4

Comparison of First and Second Tour Domains by Number of Tasks and Clusters

Occupation	Number of Tasks		Number of Dimensions	
	First Tour	Second Tour	First Tour	Second Tour
Infantryman (11B)	221[a]	246	12	11
Cannon Crewmember (13B)	177	225	11	14
Tank Crewmember (19E)	227	290	11	13
Single Channel Radio Operator (31C)	170	209	11	13
Motor Transport Operator (64C/88M)	119	150	12	12
Admin. Specialist (71L)	161	183	9	12
Medical Specialist (91A/B)	239	299	10	12

Note. The 46-item SRQ and its 8 dimensions have not been included in this table.

[a]This includes 14 tasks from a dimension which was dropped for second tour soldiers.

Table 5

New Second Tour Responsibilities by Army Job: Number of tasks in Clusters

Domain Cluster Type/Name		Occupation						
		11B	13B	19E	31C	88M	71L	91A
Technical								
Operations		22[b]	12	16	0	3	3	0
Administration		0	0	0	29	0	6[c]	33
Supervisory	N[a]							
Plan, Organize, Monitor	8	8	8	22	18	12	14	13
Clarify Roles, Provide Feedback	4	4	4	4	5	4	5	5
Provide Information	9	13	10	15	9	9	16	12
Recognize, Reward	5	5	5	5	6	5	6	7
Train Develop	8	11	17	11	11	9	11	13
Support	4	4	5	4	4	4	6	6
Discipline, Punish	5	5	5	5	5	5	5	8
Act as Role Model	3	3	3	3	3	3	3	3
Total	46	53	57	69	61	51	66	67

Note. [a]Number of tasks per SRQ category. [b]Augmented category. [c]This category was larger for first tour. Six AOSP - -ks originally in this category were reassigned to Supervisory categories.

Table 6

Importance of Supervisory Clusters and Cluster Groups Across Occupations

	OCCUPATION					
	11B	13B	19E	31C	88M	91A
Supervisory Category						
Plan, Organize, Monitor	4.07	3.66	3.78	4.00	3.90	3.42
Clarify Roles	3.25	3.58	4.10	3.98	4.09	3.10
Provide Information	4.07	3.56	3.78	4.18	4.16	3.46
Recognize, Reward	3.66	3.66	4.23	4.00	4.02	3.34
Train, Develop	3.83	3.51	4.03	4.01	4.14	3.20
Support	3.69	3.64	4.26	4.07	4.38	3.16
Discipline, Punish	3.64	3.56	4.14	4.08	4.04	2.47
Act as Model	4.65	4.41	4.48	4.31	4.34	3.87
Cluster						
Mean of Common Clusters	4.01	3.81	3.92	3.87	3.82	3.56
Mean of Technical Clusters	4.12	3.80	4.19	3.54	3.77	3.03
Mean of Supervisory Clusters	3.86	3.70	4.10	4.08	4.13	3.31

References

Ash, R. A. (1982). Job elements for task clusters: Arguments for using multi-methodological approaches and a demonstration of their utility. Public Personnel Management Journal, 11, 80-90.

Bass, B. M. (1985). Leadership and Performance beyond expectations. New York: Free Press.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1985). Development and field test of task-based MOS-specific criterion measures. (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Headquarters Department of the Army (1983). Military leadership, FM 22-100. Washington, DC: Author.

Hough, L. M., Gast, I. F., White, L. A., & McCloy, R. (1986, August). The relation of leadership and individual differences to job performance. Paper presented at the Annual Meeting of the American Psychological Association, Washington, D.C.

Pulakos, E. D., Hanson, M. A., Borman, W. C., Hallam G., Carter G. & Owens-Kurtz, C. (1987, August). Developing Behavioral rating scales to evaluate second tour performance in the Army.

Rumsey, M. G. (1987, August). Getting answers to the right questions: Job Analysis Strategy. Paper for presentation at the Annual Meeting of the American Psychological Association, New York.

Steinberg, A. G., van Rijn, P., & Hunter, F. T. (1986, November). Leader requirements task analysis. Paper presented at the 28th Annual Military Testing Association Conference, Mystic, CT.

Steinberg, A. G. (1987, May). Using task analysis to identify Army leader job requirements. Paper presented at the Sixth International Air Force Occupational Analyst Workshop, San Antonio, TX.

White, L. A., Gast, I. F., & Rumsey, M. G. (1985, November). Leader behavior and the performance of first term soldiers. Paper presented at the 27th Annual Meeting of the Military Testing Association, San Diego, CA.

White, L. M., Gast, I. F., Rumsey, M. G., & Sperling, H. M. (1986). Categories of leaders' behavior that influence the performance of enlisted soldiers. (ARI Working Paper RS-WP-86-1). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Yukl, G. A. (1981). Leadership in organizations. Englewood Cliffs, NJ: Prentice-Hall.

Yukl, G. A. (1984). Revised leader behavior task list (unpublished).

OVERCOMING OBJECTIONS TO THE USE OF TEMPERAMENT VARIABLES IN SELECTION

Leaetta M. Hough

Personnel Decisions Research Institute

Presented on Symposium,

"New Perspectives on Personality and Job Performance"

At the Annual Convention of the
American Psychological Association
New York

August 1987

Overcoming Objections to the Use of Temperament Variables in Selection

In 1982, I was assigned responsibility for developing temperament, biodata, and interest measures for Project A, a major research project funded by the Army Research Institute to improve prediction of job performance of Army enlisted personnel.

When we started, much of the scientific community believed it would be a waste of time to include temperament variables in a selection battery. There were at least five sources of negative opinion. First, in 1966 Guion and Gottier published an article in Personnel Psychology that affected the scientific community's attitude and knowledge about the usefulness of temperament variables for predicting job performance criteria. They reviewed the criterion-related validities of temperament variables and concluded that, though temperament variables have criterion-related validity more often than can be expected by chance, no generalized principles could be discerned from the results.

A second source of negative opinion about temperament variables came in the form of a theoretical challenge. In 1968, Walter Mischel published his highly influential book that caused an intense examination of and debate over trait conceptions. Mischel asserted that the apparent evidence of cross-situational consistency of behavior was a function of the use of self report as the measurement approach, that traits were an illusion. He proposed "situationism," stating that behavior is explained more by differences in situations than differences in people.

Thus, in 1982 much of the scientific community was persuaded by the published literature and believed that temperament measures had little theoretical merit and were of little practical use. Even those who

thought temperament measures might have some merit were concerned that temperament scales might be inappropriate and unfair to people who were protected under the 1964 Civil Rights Act. In addition, many people worried about intentional distortion of self descriptions in an applicant setting.

Equally important and negative was the lay community's perception of temperament inventories. People objected to offensive items and resented being asked to respond to such items. Researchers had been sensitized by the lay community's negative reaction to temperament inventories and were legitimately leery of antagonizing the public.

This was the environment in 1982.

Now, in 1987, Army generals are asking us to implement the temperament inventory we developed. What did we do to bring this about?

RESEARCH STRATEGY

A lot of time and effort was required. We also had a research strategy. That strategy is outlined on page two of your handout.* I'd like to describe that approach and some of our findings. The research strategy was construct oriented and included four basic steps: (1) a literature review to identify predictor constructs that were likely to predict job performance criteria important to the Army, (2) the development of a temperament inventory that consisted of nonsensitive items and scales designed to detect intentional distortion of self descriptions, (3) a criterion-related validity study to identify temperament scales that were job-related, and (4) an examination of the effects of motivational sets on scale scores and criterion-related validities.

* Reproduced at the end of this paper.

Literature Review

Predictor and criterion taxonomies. Since our approach was construct oriented for both predictors and criteria, we needed a taxonomy for both predictors and criteria. The criterion categories were education, training, job involvement, job proficiency, and adjustment. For the predictors, we started with the structure initially found by Tupes and Christal (1961) in the early 60s. Following Hogan's thinking in the early 80s, we split one of the constructs into two. Thus, our predictor taxonomy consisted of six constructs: Surgency, Affiliation, Adjustment, Agreeableness, Dependability, and Intellectance.

Categorization of temperament scales. Once we had a predictor taxonomy, our next step was to categorize existing temperament scales into the classification scheme. From articles and manuals, we obtained hundreds of correlations between temperament scales. We categorized the temperament scales into the six categories and a miscellaneous category, and then refined the classifications through an iterative process of classifying and reclassifying temperament scales to maximize the mean within-category correlations and minimize the mean between-category correlations. The results of this process are shown in Table 1 of your handout. The circles in the diagonal show the mean within-category correlations which are in the .30s and .40s and are, in all cases, higher than the mean between-category correlations.

Meta analysis of criterion-related validities. Our next step was to summarize the criterion-related validities according to these constructs; Table 2 of your handout shows the results. It is a meta analysis of the criterion-related validities of scales within each predictor construct for each criterion construct. As you can see, several temperament constructs correlate with the criteria. Note that there are three

additional predictor constructs. These three, "Achievement," "Masculinity," and "Locus of Control," were all a part of the miscellaneous category. When we summarized the validities for the miscellaneous category, we found respectable validities there too, so we looked more closely at the scales included in the miscellaneous category and found these additional three constructs.

The results in this table are different from the results that Guion and Gottier obtained. We believe that our strategy of summarizing the validities according to both predictor and criterion constructs accounts for the difference in results. To test this hypothesis, we summarized the validity coefficients in our database without regard to construct and obtained a coefficient of essentially zero, quite different from the coefficients in Table 2. We believe this demonstrates the importance of constructs as organizing principles for examining and understanding the literature on the criterion-related validity of temperament variables. We used the results in this table to guide us in selecting predictor constructs to measure.

Development of Temperament Scales

The next step in our research strategy was to develop measures of the constructs that the literature review indicated were likely to predict criteria important to the Army. List 1 of your handout shows the substantive scales we developed for each construct. We developed measures for six constructs: Surgency, Adjustment, Agreeableness, Dependability, Achievement, and Locus of Control. We also developed a "Physical Condition" scale and four response validity scales: Non-Random Response, Social Desirability, Poor Impression, and Self-Knowledge. We developed the Non-Random Response scale to detect inven-

tories that had been completed carelessly, a "Social Desirability" scale to detect intentional distortion that might occur in an applicant setting or a non-draft setting, and a "Poor Impression" scale to detect intentional distortion that might occur in a draft setting. We called the inventory the ABLE, short for Assessment of Background and Life Experiences.

We revised the items and scales in the ABLE many times. People representing a variety of perspectives reviewed the items for sensitive content. We also pretested the scales three times, each time evaluating and revising the items and scales based on soldiers' verbal feedback, item response distributions, internal consistency estimates, and test-retest reliabilities. The scale statistics for the ABLE scales appear in Table 3 of your handout. The average number of items in a scale is 15. The median alpha of the substantive scales is .81, and the median test-retest reliability of the substantive scales is .78. Table 4 summarizes the ABLE substantive scale statistics as well as correlations of the ABLE substantive scales with each other and with other components of the four-hour predictor battery. The only part of the predictor battery that the ABLE substantive scales correlate with in any sizable way are other ABLE substantive scales. The ABLE substantive scales appear to be tapping a part of the predictor domain not tapped by other measures.

Demonstration of Job-Relatedness

The next step in our research strategy was to demonstrate the job-relatedness of our temperament scales. We conducted a concurrent validity study during the summer and fall of 1985. Over 9000 soldiers completed the 4-hour predictor battery that included measures of cognitive

ability, spatial ability, perceptual psychomotor ability, work environment preferences, interests, and temperament.

Criterion-related validities. The criterion measures, the development of which was a major part of the research project, were developed by a different part of the research team. The criterion composites are very briefly described in List 2 of your handout. There are five composites: Core Technical Proficiency, General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing. The first two consist mainly of work samples and knowledge tests. The other three consist of supervisory and peer ratings and information obtained from personnel records.

Table 6 of your handout shows the criterion-related validities of the ABLE scales for these five criteria. The results suggest that Achievement scales are the best predictors of the "Effort and Leadership" criterion; Dependability scales are the best predictors of the "Personal Discipline" criterion; and Physical Condition is the best predictor of the "Physical Fitness and Military Bearing" criterion, though the Achievement scales also correlate with this criterion. These three criteria include the supervisory and peer ratings. The other two criteria Core Technical Proficiency and General Soldiering Proficiency, which consist of work sample and knowledge tests, are not predicted with the ABLE substantive scales.

Table 7 in your handout shows the criterion-related validities of the different types of predictors included in the study. It shows the multiple correlations of each type of predictor with each of the five criteria. As you can see, the best predictors of the supervisory and peer rating criteria, that is, Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing, are the ABLE substan-

tive scales. The other conclusion from this table is that the ASVAB mental ability test and the ABLE temperament inventory are the two best predictors of the criterion domain.

Fairness

We next turned to the issue of fairness. Are the items and scales fair for groups protected under the 1964 Civil Rights Act? The mean scores for whites, blacks, and Hispanics appear in Table 8 of your handout. As you can see, minorities do not tend to score lower than whites on the ABLE scales. Our efforts to write items that were not biased against minorities appear to have been successful. We're currently conducting differential validity and fairness analyses; those analyses, however are not yet complete.

Examination of Effects of Motivational Set

The fourth component of our research strategy involved investigating several issues related to motivational set. A frequent criticism of self-report inventories is that respondents can intentionally distort their responses. When respondents are applicants, this is an especially important criticism because the criterion-related validities might be negatively affected by distorted responses. We therefore studied the impact of motivational set on criterion-related validities, the extent to which applicants distort their self descriptions, and the usefulness of the four response validity scales to detect and adjust for motivational set.

Faking study. First, we conducted an experiment in which soldiers were instructed to respond honestly or to distort their responses in a specified way. The participants in the experiment were 245 enlisted

soldiers at Ft. Bragg. The design was a repeated measures with faking and honest conditions counter-balanced. We performed a multivariate analysis of variance on the ABLE scales and found that soldiers can distort their responses when instructed to do so.

We then examined the extent to which the response validity scales detected intentional distortion. Table 9 of your handout shows the results. The last two columns show the effect size of the difference between honest and fake good and honest and fake bad. Effect size can be interpreted in standard deviation terms. Thus, the difference in the honest and fake good condition for Social Desirability is essentially one standard deviation; the Social Desirability scale detects distortion in the fake good condition. As you can see, the Non-Random Response, Poor Impression, and Self-Knowledge scales detect distortion in the fake bad condition.

We next examined the extent to which we could use the response validity scales Social Desirability and Poor Impression to adjust ABLE substantive scales for faking. Table 10 shows the effect of regressing out Social Desirability in the fake good condition and the effect of regressing out Poor Impression in the fake bad condition. Median values are reported in this table. The .49 in the upper left-hand cell indicates that the median difference in ABLE scores between the honest and fake good condition before regressing out Social Desirability is .49 or half a standard deviation. That is, ABLE scale scores differ by about half a standard deviation in the fake good condition as compared to the honest condition. The next number to the right shows that after regressing out Social Desirability from the fake good condition, the ABLE substantive scales differ from the honest condition by only .14 or just over one-tenth of a standard deviation.

The next two values to the right show the results for the honest and fake bad conditions. Clearly, the Social Desirability and Poor Impression scales can be used to adjust substantive scale scores for intentional distortion.

These data demonstrate that: (1) people can distort their responses to temperament scales, (2) response validity scales can detect such distortion, and (3) the response validity scales can be used to adjust temperament scale scores for distortion.

We then asked, to what extent do applicants distort their responses? To answer this question, we compared scale scores of 121 Army applicants with scale scores of two groups of soldiers who had no motive for distorting their responses. Table 11 shows the results. On the substantive scales, applicants actually scored lower than one or both groups of soldiers 9 out of 11 times. These data suggest that applicants do not appear to distort their responses.

Nevertheless, we examined the effects of inaccurate self descriptions, as detected by the response validity scales, on criterion-related validities obtained in the concurrent validity study. Table 12 shows that validities for the group detected as responding in a random way are significantly lower than validities for the group responding conscientiously. Table 13 shows the increment in validity when Social Desirability is used as a moderator variable. Table 14 shows the increment in validity when Poor Impression is used with each substantive scale in a multiple correlation. The data in these three tables indicate that the response validity scales do improve, modestly, the validities of the substantive scales even in a concurrent validity study where there is little motive to distort one's self description.

Project A researchers are currently conducting a predictive validity study which will provide an opportunity to evaluate the validities of the ABLE substantive scales and the usefulness of the response validity scales in a selection situation.

Summary

We overcame objections to the use of temperament variables in selection by:

1. reviewing the literature using a construct-based approach to identify useful temperament constructs in previous criterion-related validity studies;

2. focusing scale development on constructs that are likely to predict criteria important to the client;

3. developing scales that consist of items acceptable to the public;

4. developing scales that are not biased against minorities;

5. developing scales that are psychometrically good;

6. developing response validity scales to detect inaccurate self descriptions;

7. evaluating job-relatedness of scales by demonstrating criterion-related validity;

8. developing and evaluating "adjustments" to substantive scale scores based on response validity scale scores, and;

9. evaluating the effect of motivational set on scale scores and criterion-related validities.

REFERENCES

Guion, R. M., & Gottier, R. F. (1966). Validity of personality measures in personnel selection. Personnel Psychology, 18, 135-164.

Mischel, W. (1968). Personality and assessment. New York: Wiley.

Table 1 Mean Within-Category and Between Category Correlations of Temperament Scales

	Surgency	Adjustment	Agreeableness	Dependability	Intellectance	Affiliation	Miscellaneous
Surgency	Mean$_r$=.46 SD$_r$=.16 N$_r$=146						
Adjustment	Mean$_r$=.20 SD$_r$=.18 N$_r$=321	Mean$_r$=.43 SD$_r$=.19 N$_r$=165					
Agreeableness	Mean$_r$=.04 SD$_r$=.17 N$_r$=173	Mean$_r$=.24 SD$_r$=.16 N$_r$=162	Mean$_r$=.37 SD$_r$=.14 N$_r$= 44				
Dependability	Mean$_r$=-.08 SD$_r$= .16 N$_r$= 286	Mean$_r$=.13 SD$_r$=.20 N$_r$=276	Mean$_r$=.06 SD$_r$=.17 N$_r$=166	Mean$_r$=.34 SD$_r$=.18 N$_r$=121			
Intellectance	Mean$_r$=.12 SD$_r$=.15 N$_r$=175	Mean$_r$=.02 SD$_r$=.14 N$_r$=193	Mean$_r$=.04 SD$_r$=.16 N$_r$= 94	Mean$_r$=-.12 SD$_r$= .18 N$_r$= 162	Mean$_r$=.40 SD$_r$=.19 N$_r$= 52		
Affiliation	Mean$_r$=.09 SD$_r$=.21 N$_r$=157	Mean$_r$=.00 SD$_r$=.16 N$_r$=150	Mean$_r$=.10 SD$_r$=.17 N$_r$= 98	Mean$_r$=.08 SD$_r$=.14 N$_r$=160	Mean$_r$=-.14 SD$_r$= .15 N$_r$= 84	Mean$_r$=.33 SD$_r$=.16 N$_r$= 45	
Miscellaneous	Mean$_r$=.09 SD$_r$=.17 N$_r$=392	Mean$_r$=.12 SD$_r$=.18 N$_r$=419	Mean$_r$=.02 SD$_r$=.18 N$_r$=215	Mean$_r$=.02 SD$_r$=.18 N$_r$=361	Mean$_r$=.04 SD$_r$=.17 N$_r$=242	Mean$_r$=-.04 SD$_r$= .15 N$_r$= 208	Mean$_r$=.05 SD$_r$=.20 N$_r$=246

Table 2 Meta Analysis of Criterion-Related Validity Studies [1] That Used Temperament Predictors

	Criterion											
	Educational		Training		Job Involvement		Job Proficiency		Negative Adjustment: Delinquency		Negative Adjustment: Substance Abuse	
Predictor Construct [2]	Number Predictors	mean r	Number Predictors	mean r	Number Predictors	mean r	Number Predictors	mean r	Number Predictors	mean r	Number Predictors	mean r
*Surgency	42	.15	47	.08	21	.04	175	.04	8	-.29	30	.06
Affiliation	5	-.04	0	---	4	.06	16	-.01	0	---	4	-.03
*Adjustment	44	.26	44	.16	21	.13	146	.13	10	-.43	31	-.07
*Agreeableness	9	.01	5	.10	4	.02	48	-.01	1	-.31	8	-.04
*Dependability	24	.15	26	.11	18	.17	102	.13	10	-.27	25	-.28
*Intellectance	6	.18	7	.14	8	-.10	32	.01	1	-.24	2	.19
Achievement	8	.30	4	.33	4	.24	0	---	4	-.35	0	---
Masculinity	8	-.16	3	.09	10	.10	0	---	3	.02	8	-.18
Locus of Control	1	.32	2	.29	7	.25	0	---	0	---	0	---

[1] Time Period 1960-1984.

[2] A star denotes the construct is one of the "Big Five" constructs.

Note: Correlations are not corrected for unreliability or range restrictions.

Table 3 ABLE Scale Statistics for Total Group[1]
(Concurrent Sample; Revised Trial Battery)

ABLE SUBSTANTIVE SCALES	No. Items	N	Mean	S.D.	Internal Consistency Reliability (Alpha)	Test-Retest Reliability[2]
Emotional Stability	17	8522	39.0	5.45	.81	.74
Self-Esteem	12	8472	28.4	3.70	.74	.78
Cooperativeness	18	8494	41.9	5.28	.81	.76
Conscientiousness	15	8504	35.1	4.31	.72	.74
Nondelinquency	20	8482	44.2	5.91	.81	.80
Traditional Values	11	8461	26.6	3.72	.69	.74
Work Orientation	19	8498	42.9	6.06	.84	.78
Internal Control	16	8485	38.0	5.11	.78	.69
Energy Level	21	8488	48.4	5.97	.82	.78
Dominance	12	8477	27.0	4.28	.80	.79
Physical Condition	6	8500	14.0	3.04	.84	.85
ABLE RESPONSE VALIDITY SCALES						
Social Desirability	11	8511	15.5	3.04	.63	.63
Self-Knowledge	11	8508	25.4	3.33	.65	.64
Non-Random Response[3]	8	9188	7.4	1.19	—	.30
Poor Impression	23	8492	1.5	1.85	.63	.61

[1] Total group after screening for missing data and random responding.
[2] N = 408 - 412 for test-retest correlations (N = 414 for Non-Random Response test-retest correlations).
[3] Screened only for missing data.

List 1

ABLE[1] Scales Organized According to Construct Intended to Measure

SUBSTANTIVE SCALES:

Surgency
- Dominance
- Energy Level

Adjustment
- Emotional Stability

Agreeableness (Likeability)
- Cooperativeness

Dependability
- Nondelinquency
- Traditional Values
- Conscientiousness

Achievement
- Work Orientation
- Self Esteem

Locus of Control
- Internal Control

Physical Condition
- Physical Condition

RESPONSE VALIDITY SCALES:

- Non-Random Response
- Social Desirability
- Poor Impression
- Self-Knowledge

[1] Inventory developed by PDRI for the Army Research Institute entitled "Assessment of Background and Life Experience."

Table 4 ABLE Substantive Scales: Summary

(Revised Trial Battery)

	Range	Median
Reliability:		
Internal Consistency (Alpha)	.69 - .84	.81
Test-Retest	.69 - .85	.78
Relationship to Predictor Variables:		
Correlation ABLE Substantive Scales	.00 - .73	.30
Correlation Interest Scales	.00 - .43	.09
Correlation Preferred Work Environment Scales	.00 - .35	.13
Correlation Perceptual/Psychomotor Measures	.00 - .13	.03
Correlation Cognitive Measures	.00 - .20	.05
ASVAB[1] Adj. R^2	.01 - .04	.01

[1] Mental ability test currently used by military.

Table 5

ABLE Substantive Scales: Factor Analysis[1]

(Concurrent Sample; Revised Trial Battery)

	Factor I Ascendancy	Factor II Dependability	Factor III Adjustment	h^2
Dominance	[.84]	.04	.17	.73
Self-Esteem	[.77]	.13	.37	.75
Work Orientation	[.72]	.47	.15	.77
Energy Level	[.66]	.32	.47	.76
Traditional Values	.13	[.84]	.10	.73
Nondelinquency	.04	[.82]	.26	.74
Conscientiousness	.49	[.72]	.07	.76
Internal Control	.27	.47	.44	.48
Emotional Stability	.33	.04	[.85]	.84
Cooperativeness	.17	.46	[.67]	.69
				7.25

[1] Principal component analysis, Varimax rotation.

Note: N = 8367

List 2

Criterion Composites[1]

Core Technical Proficiency - a) hands-on tests of MOS-specific technical knowledge and skills; and b) tests of school and job knowledge.

General Soldiering Proficiency - a) hands-on tests of general soldiering skill; and b) general soldiering knowledge and skill test items.

Effort & Leadership - a) supervisory and peer ratings of effort and leadership, overall effectiveness, MOS effectiveness and predicted combat effectiveness; and b) letters and certificates of commendation and other achievements.

Personal Discipline - a) supervisory and peer ratings of personal control and discipline; and b) disciplinary actions and other negative indicators in personnel files.

Physical Fitness & Military Bearing - a) supervisory and peer ratings of physical fitness and military bearing; and b) physical readiness tests.

[1]Data gathered at same time as Trial Battery was administered, i.e., summer and fall of 1985.

Table 6 Validities of ABLE Scales for Job Performance Criteria: Zero-Order Correlations (Revised Trial Battery; Concurrent Validity Study)

Predictor	Criterion: Core Technical Proficiency	General Soldiering Proficiency	Effort & Leadership	Personal Discipline	Physical Fitness & Military Bearing
Surgency:					
. Dominance	.01	.01	.15	.02	.18
Achievement:					
. Self Esteem	.02	.01	[.20]	.13	[.20]
. Work Orientation	.02	.02	[.23]	.18	[.21]
. Energy Level	.02	.02	[.22]	.14	[.25]
Adjustment:					
. Emotional Stability	.02	.02	.17	.12	.16
Agreeableness (Likeability)					
. Cooperativeness	.01	.02	.15	[.21]	.14
Dependability:					
. Traditional Values	.03	.06	.13	[.25]	.16
. Non-delinquency	.05	.07	.12	[.29]	.14
. Conscientiousness	.02	.02	.18	[.23]	.22
Others:					
. Internal Control	.04	.05	.13	.13	.13
. Physical Condition	-.04	-.05	.09	-.03	[.29]
Response Validity Scales:					
. Non-Random Response[1]	.13	.14	.07	.10	.02
. Social Desirability	-.07	-.06	.02	.05	.07
. Poor Impression	-.04	-.05	-.15	-.15	-.16
. Self-Knowledge	-.04	-.03	.07	.05	.13

[1] Correlations are based on unscreened data for this scale. N varies from 8424 to 9322 for this scale.

Note: N varies from 7666 to 8477.

Note: A box indicates notable predictor/criterion construct relationships.

Table 7

Multiple Correlations[1] of Six Independent Predictor Composites with each of Five Job Performance Criteria

(Concurrent Validity Study)

Predictor Composites	Criterion Composites: Core Technical Proficiency	General Soldiering Proficiency	Effort & Leadership	Personal Discipline	Physical Fitness & Military Bearing
ASVAB[2] (mental ability test)	**.62**	**.64**	**.35**	.20	.14
Spatial Abilities	.56	.62	.26	.14	.11
Perceptual/Psychomotor Abilities (computerized)	.54	.58	.30	.12	.10
Work Environment Preferences	.28	.27	.20	.10	.11
Temperament (and physical activities scale)	.26	.24	**.34**	**.33**	**.36**
Interests	.34	.34	.26	.14	.13

[1]Multiple Rs are adjusted for shrinkage and corrected for restriction in range, but not corrected for criterion unreliability.

[2]Mental ability test currently used by military.

Note: Entries in table are averaged across 9 Army military occupational specialties (MOS) with complete criterion data. Total sample is 3902. Sample sizes range from 281 to 570; median = 432.

Note: Boxes denote the two best predictors of the criterion space.

Table 8

ABLE Scale Means and Standard Deviations Separately for Race (Trial Battery) (Revised)

	Black (N = 2227 - 2256)		Hispanic (N = 284 - 292)		White (N = 5614 - 5673)		Other (N = 328 - 332)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ABLE Substantive Scales								
Emotional Stability	39.3	4.97	38.7	5.25	38.9	5.63	38.2	5.47
Self-Esteem	28.7	3.32	28.7	3.49	28.4	3.83	27.8	4.02
Cooperativeness	42.6	5.02	41.9	4.92	41.6	5.38	41.6	5.18
Conscientiousness	35.7	3.68	36.1	4.08	34.7	4.53	35.7	3.80
Nondelinquency	45.4	5.18	45.8	5.96	43.7	6.11	44.8	5.93
Traditional Values	27.2	3.11	27.0	3.16	26.3	3.95	26.7	3.42
Work Orientation	43.1	5.51	43.5	5.44	42.8	6.31	43.2	5.80
Internal Control	37.8	4.55	38.2	4.50	38.1	5.37	38.4	4.54
Energy Level	48.6	5.35	49.6	5.49	48.3	6.21	48.2	5.92
Dominance	27.7	3.86	27.3	4.09	26.8	4.42	26.5	4.15
Physical Condition	14.4	2.84	14.0	3.11	13.8	3.10	13.6	3.09
ABLE Response Validity Scales								
Social Desirability	15.8	3.05	17.4	3.60	15.2	2.91	17.0	3.50
Self-Knowledge	26.2	3.10	25.4	3.12	25.1	3.39	25.5	3.11
Non-Random Response	7.6	0.65	7.6	0.68	7.7	0.54	7.6	0.62
Poor Impression	1.4	1.66	1.4	1.57	1.5	1.94	1.6	1.91

Note: A box indicates a difference from the white mean of approximately one-half standard deviation or more.

Table 9

ABLE Response Validity Scales:
Effects of Honest* and Faking* Conditions
Ft. Bragg

ABLE Response Validity Scale	Honest First*			Fake Good First*			Fake Bad First*			Effect Size Honest vs. Fake Good	Effect Size Honest vs. Fake Bad
	N	M	S.D.	N	M	S.D.	N	M	S.D.		
Social Desirability (Unlikely Virtues)	109	15.8	3.1	57	20.1	5.8	56	17.8	4.8	-1.02	- .53
Self-Knowledge	109	29.6	3.6	57	29.7	4.1	56	21.8	5.2	- .03	1.85
Non-Random Response	109	7.6	1.0	57	7.0	1.8	56	2.8	2.2	.45	3.16
Poor Impression	109	1.5	2.1	57	1.7	2.2	56	14.6	7.9	- .09	-2.67

*Values are based on the sample that completed the questionnaires under the condition of interest first.

TABLE 10

Effects of Regressing Out Response Validity Scales
(Social Desirability and Poor Impression)
in Faking Conditions for ABLE

	Honest vs. Fake Good Effect Size		Honest vs. Fake Bad Effect Size	
	Before Adjustment	After Adjustment	Before Adjustment	After Adjustment
ABLE Substantive Scales	.49	.14	2.10	.45

Table 11

Comparison of Ft. Bragg Honest*, Ft. Knox, and MEPS (Applicants) ABLE Scales

ABLE Scale	Ft. Bragg (Honest)* N	Ft. Bragg (Honest)* Mean	MEPS (Applicants) N	MEPS (Applicants) Mean	Ft. Knox N	Ft. Knox Mean	Total S.D.
Response Validity Scales							
Social Desirability	116	15.91	121	16.63	276	16.60	3.21
Self-Knowledge	116	29.54	121	28.03	276	29.64	3.63
Non-Random Response	116	7.58	121	7.79	276	7.75	.64
Poor Impression	116	1.50	121	1.05	276	1.54	1.84
Substantive Scales							
Emotional Stability	112	66.22	118	66.03	272	65.05	7.86
Self-Esteem	112	34.77	118	34.04	272	35.12	5.00
Cooperativeness	112	53.33	118	54.60	272	54.19	6.05
Conscientiousness	112	46.37	118	46.49	272	48.97	5.86
Non-Delinquency	112	53.24	118	54.36	272	55.49	6.91
Traditional Values	112	36.67	118	36.97	272	37.28	4.50
Work Orientation	112	59.71	118	58.37	272	61.40	7.73
Internal Control	112	49.48	118	51.90	272	50.37	6.13
Energy Level	112	57.56	118	56.67	272	57.19	6.95
Dominance	112	35.54	118	32.84	272	35.41	6.05
Physical Condition	112	32.96	118	28.27	272	31.08	7.49

*Scores are based on persons who responded to the honest condition first.

Table 12. Moderating[1] Effects of Random Responding on Correlations Between ABLE Scales and Job Performance Criteria

ABLE SCALE	CRITERION: Effort/Leadership Low (Random)	Effort/Leadership High (Non-Random)	Personal Discipline Low (Random)	Personal Discipline High (Non-Random)	Physical Fitness/Bearing Low (Random)	Physical Fitness/Bearing High (Non-Random)
Surgency:						
Dominance	.06	.15	.05	.02	.18	.18
Achievement:						
Self-Esteem	-.00	.15	.03	.09	.08	.18
Work Orientation	.05	.23	.08	.18	.12	.21
Energy Level	.07	.22	.10	.14	.20	.25
Adjustment:						
Emotional Stability	.11	.17	.08	.12	.09	.16
Agreeableness:						
Cooperativeness	.13	.15	.17	.21	.10	.14
Dependability:						
Traditional Values	.07	.13	.19	.25	.18	.16
Nondelinquency	.09	.12	.22	.29	.14	.14
Conscientiousness	.05	.18	.11	.23	.16	.22
Others:						
Internal Control	.00	.13	.03	.13	.05	.13
Physical Condition	-.03	.09	-.00	-.03	.16	.29

N ranges from 659 to 675 for group scoring low on "Non-Random Response" scale
N ranges from 8336 to 8477 for group scoring high on "Non-Random Response" scale
Note: Statistically significant differences at $P \leq .05$ is approximately .04.

[1] We performed a split group analysis rather than a moderated regression because the variable of interest had a highly skewed distribution.

Table 13. Moderating[1] Effects of "Social Desirability" Scale on Correlations Between ABLE Scales and Job Performance Criteria

	CRITERION					
	Effort/Leadership		Personal Discipline		Physical Fitness/Bearing	
ABLE SCALE	Non-High[2]	High[3]	Non-High[2]	High[3]	Non-High[2]	High[3]
Surgency:						
Dominance	.15	.14	.00	.06	.18	.17
Achievement:						
Self-Esteem	.21	.18	.12	.12	.21	.17
Work Orientation	.25	.20	.17	.16	.22	.17
Energy Level	.23	.20	.13	.15	.27	.20
Adjustment:						
Emotional Stability	.17	.16	.11	.12	.16	.13
Agreeableness:						
Cooperativeness	.16	.13	.20	.21	.14	.12
Dependability:						
Traditional Values	.14	.11	.26	.22	.18	.11
Nondelinquency	.13	.12	.28	.29	.14	.11
Conscientiousness	.19	.14	.22	.22	.24	.14
Others:						
Internal Control	.13	.12	.12	.15	.15	.08
Physical Condition	.08	.09	-.03	-.02	.28	.29

[1] We performed a split group analysis rather than a moderated regression because the variable of interest had a highly skewed distribution.

[2] N ranges from 5896 to 5997 for group scoring Non-High on "Social Desirability" scale

[3] N ranges from 2428 to 2480 for group scoring high on "Social Desirability" scale

Note: A statistically significant difference at $p \leq .05$ is approximately .03

Table 14. Incremental Validities of ABLE Scales When "Poor Impression" Scale is Included in Predictor Equation (Linear Model)[1]

ABLE SCALE	CRITERION: Effort/Leadership r	Effort/Leadership R	Personal Discipline r	Personal Discipline R	Physical Fitness/Bearing r	Physical Fitness/Bearing R
Surgency:						
Dominance	.15	.19	.02	.15	.18	.22
Achievement:						
Self-Esteem	.20	.22	.12	.17	.20	.22
Work Orientation	.23	.25	.18	.20	.21	.23
Energy Level	.22	.22	.14	.17	.25	.26
Adjustment:						
Emotional Stability	.17	.18	.12	.16	.16	.18
Agreeableness:						
Cooperativeness	.15	.18	.21	.22	.14	.17
Dependability:						
Traditional Values	.14	.17	.25	.26	.17	.20
Nondelinquency	.13	.17	.29	.29	.14	.18
Conscientiousness	.18	.20	.23	.24	.22	.23
Others:						
Internal Control	.13	.17	.13	.17	.13	.17
Physical Condition	.09	.16	.03	.16	.29	.31

N ~ 8400

Note: A statistically significant difference at $p \leq .05$ is approximately .02

[1]A linear model was used because the zero-order correlations of the "Poor Impression" scale with the criteria are approximately -.15.

SOURCES OF NEGATIVE OPINION

- Guion & Gottier literature review--conclude temperament measures are of little practical use
- Theoretical challenge--situationism (Mischel)
- Inappropriate and unfair for persons protected by 1964 Civil Rights Act
- Intentional distortion of self reports in applicant setting
- Offensive item content

RESEARCH STRATEGY: CONSTRUCT ORIENTATION

1. Review Literature

   - Develop predictor taxonomy
   - Classify temperament scales
   - Develop criterion taxonomy
   - Summarize criterion-related validities according to predictor and criterion constructs
   - Identify useful predictor constructs

RESEARCH STRATEGY: CONSTRUCT ORIENTATION

2. Develop Temperament Scales

- Examine items for sensitive content
- Develop response validity scales to detect intentional distortion
- Pretest
- Examine psychometric characteristics
- Revise

RESEARCH STRATEGY: CONSTRUCT ORIENTATION

3. **Demonstrate Job-Relatedness**

- **Conduct concurrent validity study**
- **Compute criterion-related validities**
- **Conduct differential validity analyses**
- **Conduct fairness analyses**
- **Conduct predictive validity study**

RESEARCH STRATEGY: CONSTRUCT ORIENTATION

4. **Examine Effects of Motivational Set**

   - **Evaluate fakability of scales**
   - **Evaluate response validity scales**
   - **Evaluate moderator effects of response validity scales**
   - **Develop "adjustment" formula**
   - **Assess effects on criterion-related validities**

SUMMARY

1. Review the literature using a construct-based approach to demonstrate the usefulness of temperament variables in previous research.

2. Focus scale development on constructs that are likely to predict criteria important to the client.

3. Develop scales consisting of items acceptable to the public.

4. Develop scales that are not biased against minorities.

5. Develop scales that are psychometrically good.

6. Develop response validity scales to detect inaccurate self descriptions.

7. Evaluate job-relatedness of scales by demonstrating criterion-related validity.

8. Develop and evaluate "adjustments" to substantive scale scores based on response validity scale scores.

9. Evaluate effect of motivational set on scale scores and criterion-related validities.

OPTIMAL JOB ASSIGNMENT AND THE UTILITY OF PERFORMANCE: SOME KEY ISSUES

Roy Nord
Leonard A. White

U.S. Army Research Institute

Presented at the Annual Convention of the
American Psychological Association
New York

August 1987

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

OPTIMAL JOB ASSIGNMENT AND THE UTILITY OF PERFORMANCE:

Some Key Issues

BACKGROUND

Decision-makers in the military services are frequently faced with policy alternatives that will produce different distributions of soldier competence across a large set of jobs. These alternatives include not only policies dealing directly with the selection, allocation and training of soldiers, but also a range of actions that affect the share of scarce resources devoted to manpower as opposed to other "inputs" to the process of producing national defense. Whenever such alternatives are compared, a judgement, either implicit or explicit, must be made as to the "value" (and cost) of different distributions of soldier performance.

The primary concern of this paper is with the measurement and use of performance utility as an aid to personnel classification and assignment decisions. Historically, most of the research on performance utility in industrial psychology has addressed the problem of translating performance gains from various selection strategies into a metric that can be used to demonstrate the value of improved selection procedures to skeptical decision-makers. The most common metric is dollar value (see, e.g., Brogden, 1959; Cascio, 1987; Hunter and Schmidt, 1982), although metrics other than dollar value have also been used (Eaton, Wing, and Mitchell, 1985).

The appropriate metric for comparisons among different classification and assignment procedures may be quite different from the metric needed to assess

selection strategies. Sadacca and Campbell (1985) noted that, in the context of optimal assignment, a dollar metric is not required and, in some cases, may be inappropriate. The reason for this difference is straightforward: In the context of selection, an organization must consider the tradeoffs between the expenditure of scarce resources (often dollars) to increase the quality of its manpower versus the expenditure of those resources to increase the quantity or quality of some other factor of production. In the case of classification and assignment, the objective is to maximize the efficiency with which a given pool of available manpower is used. In either case, the organization is concerned with the efficient allocation of a scarce resource among competing activities, but in the case of selection, manpower is an "activity" and in the case of assignment, it is a "resource".

Performance "Utility" vs Performance "Value"

Before pursuing the discussion further, a brief semantic digression is in order. The term utility has been widely used in personnel psychology to refer to what this paper will call performance value.

A reasonable argument can be made that "utility", in its most general sense, is a more appropriate term than "value", given the unquantifiable nature of the "outputs" we are analyzing. We shall employ the term "value", however for two reasons: First, to avoid confusion between our focus on the role of performance value in determining the optimal distribution of people to jobs and the more common use of performance utility in evaluating the benefits of selection and classification systems; second, to distinguish our interpretation of subjective judgements of performance value from the way

similar judgements are used in most applications of multi-attribute utility theory — which seek to identify the parameters of individual "utility functions".

In these applications, the approach we follow — that is, the averaging of individual judgements to obtain a single performance value function, would require the use of interpersonal comparisons of utility that are prohibited by the theory upon which multi-attribute utility is based (e.g., Keeney and Raiffa, 1976). This restriction does not apply to the analysis described here because we do not treat individual judgements of performance value as "utility functions" but rather as imperfect (but randomly distributed) estimates of a single organizational "value function".

The Army's Project A: Measuring the Value of Job Performance

The data reported here were collected as part of a nine-year Army research effort (Project A) aimed at improving the Army's selection and classification system for enlisted personnel. The main purpose of the utility-measurement component of Project A is to provide the information needed to maximize the payoff to the Army of improved selection and classification procedures.

The Army research on performance value assessment is being carried out in two stages. The first stage, completed earlier this year, focused on the estimation of MOS-specific performance value functions for 276 entry-level Army MOS (Sadacca, Campbell, Wise, White, 1987). In the second stage which has just begun, the resulting functions, under different configurations of constraints, will be used to produce distributions of performance across 19

selected jobs, and senior Army officers will be asked to evaluate the resulting distributions. In effect, the purpose of the second stage is to evaluate alternative strategies for implementing performance value estimates in an optimal job assignment system.

In this section, we will briefly summarize the data collection approach and results of the first stage. The following section will examine the effects of using the resulting value functions to make job assignment decisions.

The first stage consisted of a series of 7 workshops at which we obtained judgements by 74 field-grade officers of the relative value of performance at five levels in all entry-level Army MOS. At each workshop, the performance level/job combinations were scaled using two methods. In the first, each officer received one of seven decks of cards. Each card specified a performance percentile and an MOS, the duties of which were briefly described on the card. The officers were asked to sort the cards into six groups — one group representing combinations with a negative value and five representing ordinal rankings of increasingly valuable combinations. (An additional and infrequently used category was provided for combinations which could not be judged.) The seven decks combined spanned the entire set of performance level/MOS combinations for 276 MOS. In addition, each deck contained 60 combinations (12 MOS x 5 performance levels) that were common across all decks. In the second exercise, judges provided interval-level estimates of the relative value of these 60 combinations. In this exercise, the value of a 90th percentile infantryman (MOS 11B) was fixed at 100. The officers were

then asked to scale the remaining combinations relative to this fixed value. Negative values were allowed.

The performance levels were set at the 10th, 30th, 50th, 70th and 90th percentiles, using the current (1986) recruit pool as the reference population. The instructions specified that the judgements should be made under an assumption that the world was in a state of "heightened tensions". Care was taken to explain that the "performance" being evaluated was multi-dimensional, consisting not only of technical proficiency, but also personal discipline and willingness to work.

The sample of officers providing the judgements represented a cross-section of specialties. A primary consideration in this exercise was to insure that performance value judgements, to the maximum extent possible, reflected the payoffs of performance to the Army. To accomplish this, we used as judges experienced senior officers with a broad perspective on those needs. Furthermore, the effect of specialty on the judgements of performance value was generally insignificant. The judges' mean utilities had a reliability of .990 on the interval scale judgements, and from .958 to .976 across the six decks for the ordinal judgements.

To insure that the performance value estimates for all 1380 performance-level/MOS combinations were comparable, functions were estimated for each deck to transform the pile placement judgements to the interval scale used for the 60 common combinations. Table 1 contains the estimated coefficients and R^2 statistics for these functions. The robustness of the transformation functions was checked by estimating the functions using 40 c inations

selected from the set of 60, and then regressing predicted against actual values for the 20 omitted combinations. The resulting r-square statistic was .945, suggesting that the transformations should yield highly accurate estimates of the values that would have been obtained by directly scaling all 1380 combinations.

Table 1
Deck-Specific Regression Equations for Transforming Average Pile Placement to Common Interval Scale (Regressions for 60 Common Combinations)

Independent Variable	Deck A	B	C	D	E	F
Pile Placement (PP)	14.00	21.81	43.39	24.09	46.99	49.45
PP^2	1.455	-.3227	-5.785	-1.344	-6.922	-6.932
PP^3	-.0529	.0671	.4853	.1692	.5450	.5487
Intercept	-34.09	-41.75	-69.54	-47.74	-63.80	-77.85
Adjusted R^2	.965	.926	.954	.944	.912	.924

The average interval scale values at each performance level were then used to fit a "performance value" function for each job. The functions were fitted using stepwise ordinary least squares where the independent variables were performance level, its square, and its cube. The graphs of these functions for nine MOS are presented in Figure 1. These functions illustrate several interesting aspects of our results so far.

First, for most MOS, the relationship between performance level and performance value is a concave function. That is, the functions demonstrate diminishing payoffs to increases in performance as the preformance level

increases. As we shall see in the next section, this characteristic of the value functions plays a critical role in the context of optimal assignment.

A second finding is that there is substantial variety in the shape as well as the intercept (or "scale") of the functions across MOS. One can interpret the scale differences as variations in the "average" value of performance across jobs. In economic terms, this variation can be interpreted as variation across jobs in the marginal product of job output — that is, differences in the rate at which changes in productivity within a single job contribute to total Army output. Differences in the shape of the functions reflect variations in the way soldier performance at different levels contributes to job output. One would expect, for instance, that functions that are relatively "steep" at low performance levels would be associated with jobs in which the cost of errors is high; and that jobs with relatively steep slopes at high levels of performance would be those in which the payoffs to exceptional performance are relatively high.

On the other hand, as one might expect from previous work in the area of utility generalization there also appear to be identifiable groups of MOS with virtually identical functions. The task of identifying these groups and examining their similarities with respect to both the definition and prediction of performance is an important subject for further research (cf. Bobko, Karren and Kerkar, 1987).

PERFORMANCE VALUE IN CLASSIFICATION DECISIONS

In this section we address several issues associated with the use of performance value information to make manpower allocation decisions. First,

we examine the consequences of allocating people to jobs so as to maximize the value of performance, rather than performance itself. Second, we explore some of the issues associated with aggregation of performance value across many assignments. Finally, in the concluding section we raise the issue of how to determine whether or not the use of performance value information will yield better results than would be obtained without it.

The Role of Performance Value in Manpower Allocation

In general, a policy that maximizes predicted performance and ignores the value of performance will produce an allocation that has the following characteristics:

(a) the average level of performance will be highly variable across jobs;

(b) the level of expected performance will tend to be high in those jobs for which performance is easiest to measure and predict;

(c) neither job-specific differences in the way manpower contributes to output nor variations in the importance to the organization of the output from different jobs will be reflected in the allocation.

The extent to which these conditions are evident in practice will depend on the following factors:

(1) the distribution of the performance predictors in the population;

(2) the degree to which performance is differently defined in different jobs (that is, the dimensionality of performance);

(3) the variability in validities across jobs and the relationship between validity and job quotas;

(4) the extent to which the allocation is constrained by considerations other than performance.

The effects of (1) and (2) are easiest to explain if we examine them together. If we look at the extremes of the range of these two factors, two effects become clear: If performance is single-dimensioned, or if the predictors of job performance are perfectly correlated in the population, the allocation produced by maximizing expected performance will be exclusively determined by variations across jobs in the predictability of performance. If such variations do no exist, then there will be a multiplicity of equivalent "optimal" allocations. At the other extreme, if performance is uniquely defined for every job, and the predictors of performance are perfectly negatively correlated, then the allocation resulting from performance maximization will be unique and identical to the result that would be produced by maximizing any increasing function of performance. In other words, performance value will be irrelevant to the allocation problem.

With respect to the interaction between validities and job quotas noted in (3), it is obvious that the consequences of variation in predictability will become less pronounced as the variability decreases. Perhaps less obvious is the fact that, if high validities are associated with jobs that have large quotas, the effect of relatively small variations in validity can be exaggerated far out of proportion to the degree of variation.

Finally, the effect of exogenous constraints (4) is to narrow the range of feasible allocations. The more confining these constraints become, the smaller will be the difference between the "best" and "worst" feasible

allocations and thus the smaller the difference induced by considerations of either predicted performance or performance value. This factor is of particular importance in the case of the Army's allocation problem, which is circumscribed by an extensive set of policy and managerial constraints. These include not only limitations imposed by force structure requirements and the availability of training resources, but also a number of policy constraints whose purpose is to insure an acceptable, if not optimal distribution of performance across jobs. This latter set of constraints includes minimum job entry standards, an MOS priority system, and a set of job-specific "quality goals" based on educational attainment and scores on the Armed Forces Qualification Test (AFQT). One of the effects of these constraints, when they are used in optimal assignment, is to mitigate the effects of variation in validity and job quotas — producing an allocation in which average performance is lower, but also less variable across jobs than would occur without them.

Figure 2 demonstrates some of these effects for a sample of Army MOS. The distribution displayed here was produced by assigning a random sample of 2232 1984 recruits to the nine jobs so as to maximize expected performance while meeting job demands (scaled to the sample size in proportion to actual 1984 requirements) with recruits who met the minumum entry standards for the jobs to which they were assigned. The optimization used a simple network assignment algorithm that maximized the sum across all assignments of predicted performance. Predicted performance was calculated using estimated validities of current Aptitude Area scores against technical job performance.

Each bar in Figure 2 represents the mean performance level of the recruits assigned to that job, with performance level measured percentiles based on sample scores. The validity associated with each job is indicated at the tip of the bar for that job, and the sample N's are listed at the bottom of the graph.

The effects discussed above are well illustrated by these results. The distribution is highly variable across jobs. An ordering of the jobs by their validities would yield a nearly identical list to that produced by ranking average performance levels. The sole exceptions to this rule are MOS C and F. (These two jobs use different predictors, and are significantly different in size.) Finally, the interaction effect between validity and job quotas can be seen by comparing the allocation to MOS A to that for MOS C. MOS A, with a validity of .66 and a quota of 691, is assigned recruits performing, on average, at about the 70th percentile. MOS C, which uses the same predictor, has a validity only .07 less, but a quota only one seventh as large, and receives an allocation performing nearly 50 percentiles lower.

Aggregating Performance Value Across Assignments

The question we address in this section is that of using the information obtained in the exercises described above to measure the aggregate value of performance. As we shall see, the choice of an approach to this issue will have a significant effect on the distributions produced when the value functions are used in an optimal assignment algorithm.

The allocation problem can be simply described as follows:
Let N be the total number of positions to be filled, M be the number of jobs,

and K the number of levels of performance. We can then represent any assignment of N individuals to M jobs by an M x K matrix Q, where q_{ij} is the number of individuals at performance level j assigned to job i. If we define a k x 1 vector p such that p_i is the quantity of performance obtained from an individual performing at level i (the elements of p might be performance percentiles, for instance), then we can define a scalar Z, the total quantity of performance represented by the allocation Q as —

$$Z=p'Q \qquad (1)$$

That is, the total quantity of performance represented by the allocation Q is simply the sum of the number of individuals assigned to each job, weighted by performance level. This is the definition of aggregate performance that we will use. However, before continuing, it should be noted that such a definition implicitly assumes that the total quantity of performance obtained is independent of how performance is distributed within and across jobs. In other words, we are ignoring issues relating to unit or group performance.

Given this definition of aggregate performance, we must define a way of applying a performance value function to the quantity Z; that is, we must define a function v(Z) using the information obtained in the value assessment exercises described above. We shall consider two alternatives:

(a) That v(Z) is a "strongly separable" function of p and Q that can be written in the form —

$$v(Z) = \sum_{i=1}^{K} \sum_{j=1}^{M} q_{ij}\, u_j(p_i) \qquad (2)$$

where $u_j(p_i)$ is the value of performance at level i in job j.

If we assume strong separability, the marginal change in the value of performance with respect to a change in the number of individuals at a given performance level in a given job is constant, no matter how we specify the function u(p). That is,

$$\frac{\delta v(Z)}{\delta q_{ij}} = u_j(p_i), \text{ for } 0 \leq q_{ij} \leq N, \; i \varepsilon K, \; j \varepsilon M. \qquad (3)$$

(b) That v(Z) is "weakly separable" — that is

$$v(Z) = \sum_{j=1}^{M} u_j(q_j p), \text{ where } q_j \text{ is the } j^{th} \text{ row of } Q. \qquad (4)$$

By relaxing the separability assumption, we allow the marginal value of an additional assignment to a given job to vary with the total quantity of performance in that job as well as with the performance level of the particular assignment being considered:

$$\frac{\delta v(Z)}{\delta q_{ij}} = h_j(p'q') \qquad (5)$$

The consequences of this difference for optimal assignment are easily seen by comparing the maximization problems associated with the two specifications.

Let d_j represent the demand (quota) for job j, and s_i be the supply of applicants (recruits) predicted to perform at level i. (For now, we assume that performance is unidimensional — that is, each applicant will perform at the same level in all jobs.) Then the performance value function defined by (2) and (3), produce the following optimal assignment problems:

$$\text{(a)} \quad \text{Maximize} \sum_{i=1}^{K} \sum_{j=1}^{M} q_{ij} u_j(p_i) \tag{6}$$

or

$$\text{(b)} \quad \text{Maximize} \sum_{j=1}^{M} u_j(q_j p) \tag{7}$$

$$\text{Subject to:} \sum_{i=1}^{K} q_{ij} = d_j, \text{ for all } j \in M \quad \text{(demands)} \tag{8}$$

$$\sum_{j=1}^{M} q_{ij} = s_i, \text{ for all } i \in K \quad \text{(supplies)} \tag{9}$$

The equation systems defined by (a) and (b) can be transformed into single equations using the method of Lagrange as follows:

$$\text{(a)} \quad L = \sum_{i=1}^{K} \sum_{j=1}^{M} q_{ij} u_j(p_i) + \sum_{j=1}^{M} \gamma_j \sum_{i=1}^{K} q_{ij} - d_j + \sum_{i=1}^{K} \pi_i \sum_{j=1}^{M} q_{ij} - s_i \tag{10}$$

or

$$\text{(b)} \quad L = \sum_{i=1}^{K} \sum_{j=1}^{M} u_j(p' q'_j) + \sum_{i=1}^{M} \gamma_{jj} \sum_{i=1}^{K} q_{ij} - d_j + \sum_{i=1}^{K} \pi_{ij} \sum_{j=1}^{M} q_{ij} - s_i \tag{11}$$

where γ_j and π_i are sets of Lagrangian multipliers associated with the demand and supply constraints.

The conditions for a maximum of (a) will be easier to describe if we order the values of $u_j(p_i)$ so that the following is true:

$$\text{If } j' > j \text{ then } u_{j'}(p_i) \leq u_j(p_i)$$
$$\underline{\text{and}} \text{ if } i' > i \text{ then } u_j(p_{i'}) \leq u_j(p_i)$$

Then the matrix of assignments Q^* that maximizes (a) will contain elements q_{ij} that meet the following condition:

$$q_{ij} = \text{MAX}\left\{s_i - \sum_{k=1}^{i-1} q_{kj},\ d_j - \sum_{m=1}^{j-1} q_{im}\right\}. \qquad (12)$$

In other words, the maximum will be achieved by following the simple rule of "top-down" assignment: Order the set of possible person-job matches from those with the highest value to those with the lowest; then assign individuals at the highest available level of performance to the position with the highest value at that level of performance until either the demand is met or the supply is exhausted. The resulting allocation will be the one that maximizes the variance in performance value across jobs.

The necessary (first order) conditions for a Q^* that maximizes (b), the weakly separable case, can be stated as follows:

$Q^* = \{q_{ij}^*\}$ such that

$$\text{(i)} \quad \left.\frac{\delta L}{\delta q_{ij}}\right|_{q_{ij}\varepsilon q_j^*} = \left.\frac{\delta u_j(q_j^* p)}{\delta q_{ij}}\right|_{q_{ij}\varepsilon q_j^*} - \gamma_i - \pi_j = 0, \text{ for all } i,j \qquad (13)$$

$$\text{(ii)} \quad \left.\frac{\delta L}{\delta \gamma_i}\right|_{q_{ij}\varepsilon q_j^*} = s_i - \sum_{j=1}^{M} q_{ij} = 0, \text{ for all } i \qquad (14)$$

$$\text{(iii)} \quad \left.\frac{\delta L}{\delta \pi_j}\right|_{q_{ij}\varepsilon q_j^*} = d_j - \sum_{i=1}^{K} q_{ij} = 0, \text{ for all } j \qquad (15)$$

the solution of this system implies that, if the functions u_j are continuous, everywhere twice differentiable, and convex, there will exist a unique optimal solution that is characterized by the following:

$$\frac{\delta u_j(\ )/\delta q_{ij}}{\delta u_k(\ )/\delta q_{ij}} = \frac{\delta u_j(\ )/\delta q_{mj}}{\delta u_k(\ )/\delta q_{mk}} \text{ for all } i \neq m,\ j \neq k, \qquad (16)$$

and

$$\frac{\delta u_j(\ )/\delta q_{ij}}{\delta u_j(\ )/\delta q_{mj}} = \frac{\delta u_k(\ )/\delta q_{ik}}{\delta u_k(\ )/\delta q_{mk}} \text{ for all } i \neq m,\ j \neq k. \qquad (17)$$

That is, at optimality, the marginal rates of substitution across jobs for the same performance level will be the same for all pairs of jobs and performance levels, as will the marginal rates of substitution among different performance levels within jobs.

If it is reasonable to assume that the judgements obtained in the Project A utility workshops are valid, at least over a limited range, then the generally curvilinear functions displayed in Figure 1 will, under weak

separability, produce an optimal allocation that is not a "corner solution" -- that is the maximization of performance value will tend to allocate some high-level performers to all jobs. This will occur because the variations in marginal value implied by the non-constant slopes of the curves will tend to produce the equalities in (16) and (17) at values of q_{ij} that are neither 0 nor maximal. The result is that, even in MOS with very steep average slopes, there will exist some point at which the payoff to an additional assignment in this MOS is exceeded by that in another MOS with a lower average performance value.

Figures 3 and 4 illustrate the effects of weak versus strong separability on the distribution of performance for the same sample represented in Figure 2. All three figures were obtained using the same supplies, demands and constraints. The only differences are in the objective functions that were maximized.

Figure 3 presents the distribution produced when strongly separable value functions are used to maximize the aggregate value of performance. Comparing this result to Figure 2, we can see that the variability of the distribution is increased by the consideration of performance value, although the variation is less tightly linked to differences in validity. The effect of the interaction between validity and quota, however, is markedly reduced in the case of MOS A and C. The difference between MOS B and H, however, is exaggerated by the use of utility. This occurs because MOS B has both a low validity and a relatively flat value function, while the reverse is true for MOS H.

Figure 4 displays the results when weak separability is assumed. Inter-job variability is markedly reduced, with the most noticable difference being that MOS B receives a substantially increased level of performance.

Summary and Conclusions

The results pictured in Figures 2-4 provide ample evidence that variations in performance value can make substantial differences in manpower distributions. The question we shall briefly address in this section is that of determining whether or not a given approach will yield results that are "better" than the results produced by alternative approaches.

A general argument can be made to the effect that:

(1) The evidence of current practice, augmented by data from preliminary workshops to assess different distributions strongly suggests that performance value must be considered in the assignment process.

(2) the procedures currently used to insure acceptable distributions may, in the current environment, produce results that are quite good, but these procedures have three flaws (a) they are based on predictors of performance rather than predicted performance; (b) they are slow to adapt to changes in Army needs and the operating environment; and (c) they are difficult to explain and/or defend to interested parties such as DOD and Congress.

(3) Carefully estimated value functions are useful because they are (a) more adaptable to changing circumstances (in doctrine, technology, recruiting environment, etc) than are the heuristics and political mechanisms currently used to control distributions; and (b) more "rational", and thus easier to describe to others.

References

Bobko, P., Karren, R., & Kerkar, S. P. (1987). Systematic research needs for understanding supervisory-based estimates of SDy in utility analyses. Organizational Behavior and Human Decision Processes, 69, 69-95.

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, 9, 181-190.

Cascio, W. F. (1987). Costing human resources: The financial impact of behavior in organizations (2nd ed.). Boston: Kent.

Eaton, N. K., Wing, H., & Mitchell, K. J. (1985). Alternate methods of estimating the dollar value of performance. Personnel Psychology, 38, 27-40.

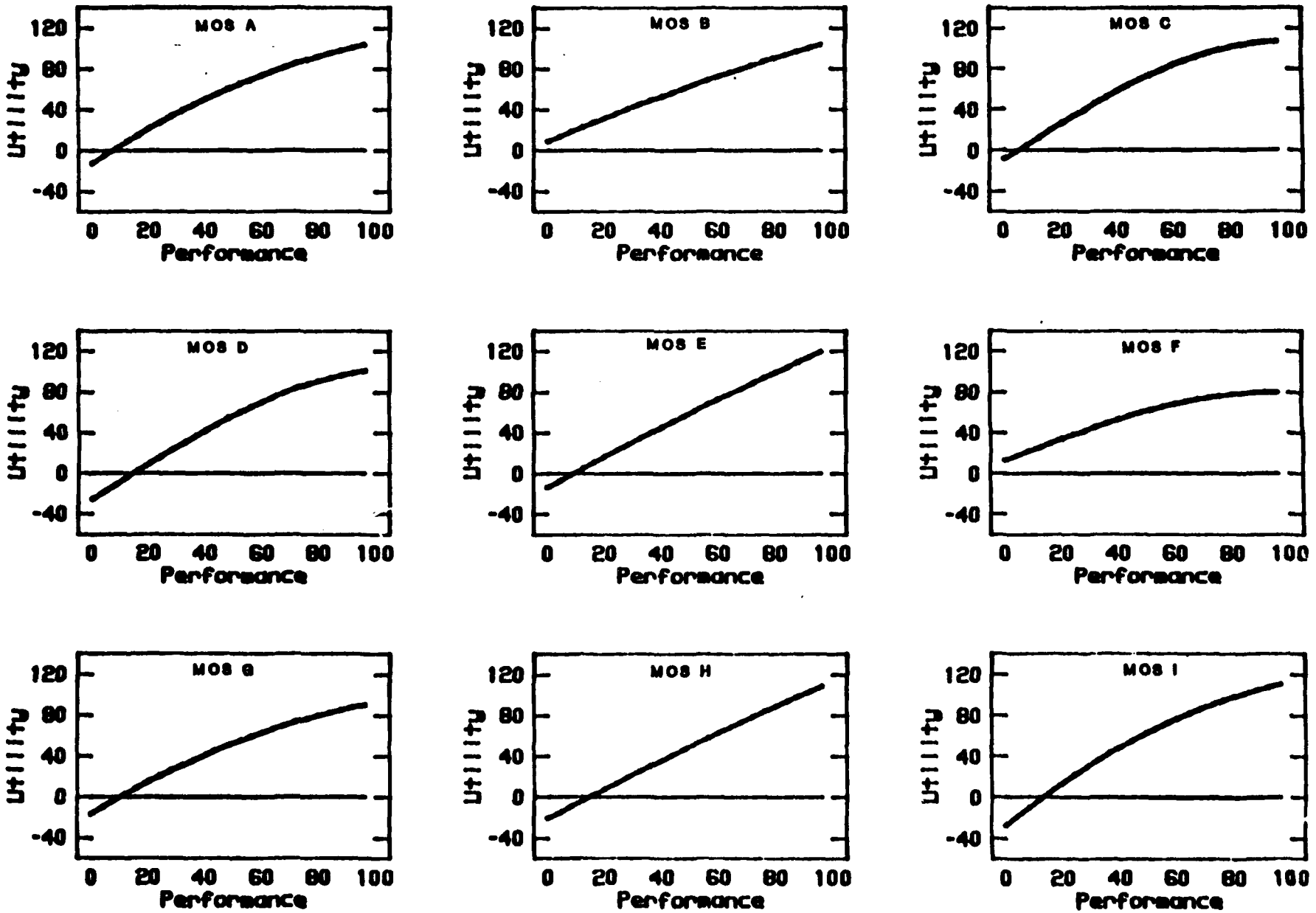
Hunter, J. E. & Schmidt, F. (1982). Fitting people to jobs: the impact of personnel selection on national productivity. In Human Performance and Productivity: Human Capability Assessment. Hillsdale New Jersey: Erlbaum.

Keeney, R. L. & Raiffa, H. (1976). Decisions with multiple objectives: Preferences and value tradeoffs. New York: Wiley.

Sadacca, R., Campbell, J., Wise, L., & White, L. (April, 1987). Performance composites, performance utility, and selection/classification decisions. Paper Presented at the second annual conference of the Society of Industrial and Organizational Psychology. Atlanta, GA.

Schmidt, F., & Hunter, J. E., Outerbridge, A. N., & Trattner, M. H. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the Federal work force: An empirically-based demonstration. Personnel Psychology, 39, 1-29.

Figure 1
Performance Utility Functions for Nine Army MOS



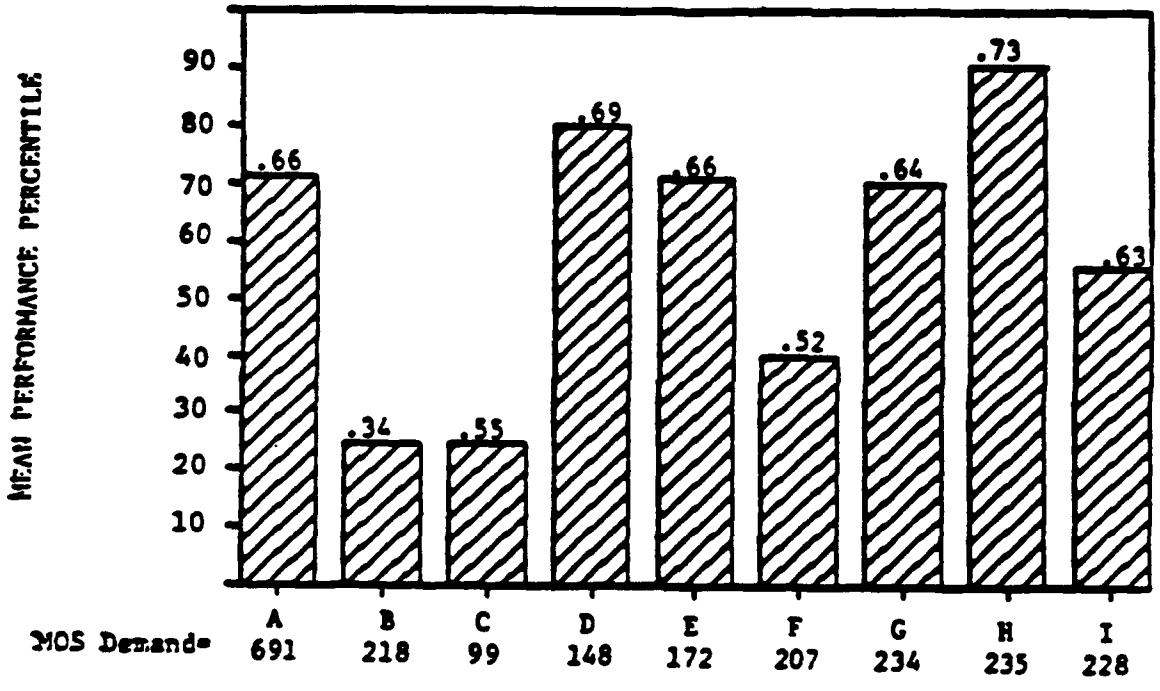


Figure 2. Expected mean performance resulting from assignment to maximize performance. Validity coefficients in each MOS are shown above the bar.

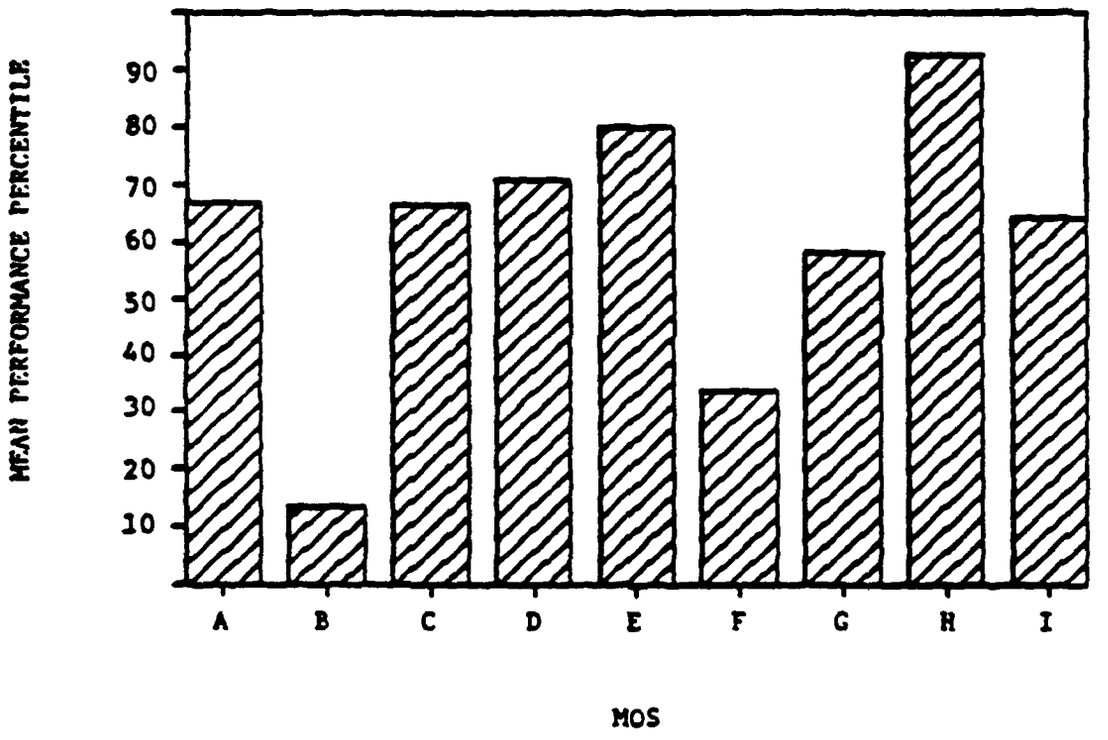


Figure 3. Expected mean performance resulting from assignment to maximize the value of performance assuming strongly separable value functions.

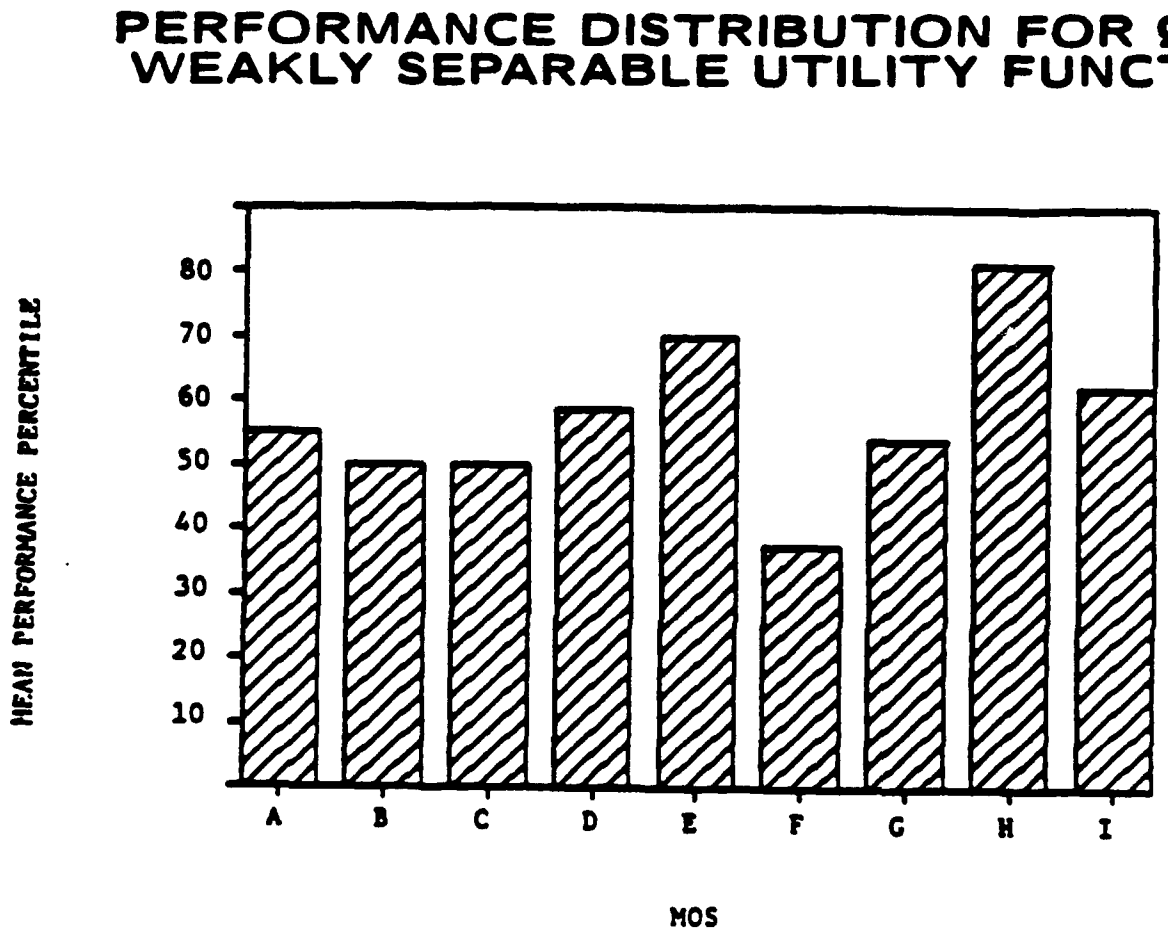


Figure 4. Expected mean performance resulting from assignment to maximize the aggregate value of performance assuming weakly separable value functions.

DEVELOPING BEHAVIORAL RATING SCALES TO EVALUATE SECOND-TOUR PERFORMANCE IN THE ARMY

Elaine D. Pulakos
Mary Ann Hanson
Walter C. Borman
Glenn Hallam
Gary Carter
Cynthia Owens-Kurtz
Personnel Decisions Research Institute

Presented on Symposium,

"Junior Noncommissioned Officer Job Requirements:
Where Does Leadership Fit In?"

At the Annual Convention of the
American Psychological Association
New York

August 1987

Developing Behavioral Rating Scales to Evaluate Second Tour Performance in the Army

Abstract

Using the critical incidents or behavioral analyses method, Army-wide and MOS-specific performance requirements were identified for second tour U.S. Army soldiers in nine representative occupational specialties. Based on these performance incidents, behavioral categories or dimensions were developed for evaluating second tour performance effectiveness. Results of both the MOS-specific and Army-wide scale development processes suggested that second tour soldiers perform most of the work that first tour soldiers perform and also supervise that work. Discussion focuses on the shortened set of procedures used to revise first tour MOS-specific rating scales to measure second tour performance and the nature of these first-line supervisor jobs in relation to the importance of technical and supervisory duties for performing effectively.

Developing Behavioral Rating Scales to Evaluate Second Tour Performance in the Army

This paper describes the procedures used to develop MOS-specific and Army-wide behavioral rating scales for evaluating the performance of second tour Army enlisted personnel. These scales were developed as part of the Project A effort to evaluate the validity of current and new predictors of soldier performance. The Project A research is being conducted on 19 Army jobs (Military Occupational Specialties or MOS), carefully selected to be representative of the entire population of Army MOS.

A primary goal of Project A is to increase Army organizational effectiveness by improving the job-soldier match. As an important step towards this goal, a comprehensive set of selection and classification measures (predictors) was developed and extensively field tested (Peterson, 1986). In addition, a diverse and comprehensive first tour criterion development effort was undertaken. Time and cost limitations dictated that job-specific criterion measures be developed for just nine of the target MOS. These job-specific criterion measures included hands-on, task proficiency measures, job knowledge tests (Campbell, Campbell, Rumsey, & Edwards, 1986), and MOS-specific behavior-based rating scales (Toquam, McHenry, Corpe, Rose, Lammlein, Kemery, Borman, Mendel, & Bosshardt, 1986). To provide criterion measurement for the remaining 10 jobs, Army-wide rating scales applicable for evaluating first tour soldier effectiveness in any MOS were developed (Borman, Motowidlo, Rose, & Hanser, 1987; Pulakos & Borman, 1986).

During the summer and fall of 1985, a large-scale Concurrent Validation was conducted, during which the predictor and criterion

measures were administered to several thousand soldiers in 19 target jobs. Then, starting in the late summer of 1986, a Longitudinal Validation data collection was begun in which all measures are being administered to approximately 50,000 soldiers in a predictive validity design. In addition to validating the predictors against first tour job performance, the measures will also be validated against second tour job performance, for those individuals in the sample who reenlist in the Army.

In this paper, we describe the procedures used to develop Behavioral Summary Scales (Borman, 1979) for evaluating second tour performance. Performance requirements for second tour U.S. Army soldiers were a priori thought to include both technical and supervisory components. That is, we believed that second tour soldiers were responsible for performing most of the technical work required of first tour soldiers and also supervising that work. Accordingly, technical competence dimensions as well as supervisory effectiveness dimensions would likely have to be incorporated into the second tour performance measures. However, the extent to which supervision is an important part of the second tour soldier's job was thought to vary across the different MOS. This suggested the possibility that some supervisory measures might be Army-wide and thus applicable to all MOS, while other supervisory measures might be MOS-specific and thus only relevant to a particular job.

Development of Second-Tour Army-Wide Rating Scales

Method and Results

Behavior Analysis Workshops

Sample. One-hundred and seventy-two officers and NCOs participated in half-day workshops intended to elicit behavioral examples of second-tour soldier effectiveness. The workshops were conducted at Ft. Bragg,

NC and Ft. Carson, CO. The sample consisted of 154 males and 18 females; 136 were officers and 36 were NCOs. These individuals reported having an average of 6.29 years in the Army and an average of 5.09 years occupying supervisory positions.

Procedure. The inductive behavioral analysis strategy (Campbell, Dunnette, Arvey, & Hellervick, 1973) requires persons familiar with a job's performance demands to generate examples of effective, mid-range, and ineffective behavior observed on that job. In the present application, job behavior was defined broadly as any action related to soldier effectiveness, and workshops were conducted in which participants were asked to generate behavioral examples of what they considered to be the second-tour soldier effectiveness domain.

A total of 1,000 behavioral examples were generated from the workshops. These incidents were edited to a common format and then content analyzed to form 12 preliminary dimensions of second tour Army-wide effectiveness. The performance categories that had been developed for the first tour soldiers were replicated for the second tour soldiers. In addition, three generic supervisory dimensions emerged from the content analysis of the incidents. Thus, categorization of the performance examples suggested that second tour soldiers do, in fact, perform most of the work that first tour soldiers perform and also supervise that work. The 12 second tour performance dimensions were as follows:

A. Displaying Technical Knowledge/Skill

B. Displaying Effort, Conscientiousness, and Responsibility

C. Organizing, Supervising, Monitoring, and Correcting Subordinates (supervisory dimension)

D. Training and Developing (supervisory dimension)

E. Showing Consideration and Concern for Subor-- dinates (supervisory dimension)

F. Following Regulations/Orders and Displaying Proper Respect for Authority

G. Maintaining Own Equipment

H. Displaying Honesty and Integrity

I. Maintaining Proper Physical Fitness

J. Developing Own Job/Soldiering Skills

K. Maintaining Proper Military Appearance

L. Controlling Own Behavior Related to Personal Finances, Drugs/Alcohol, and Aggressive Acts

Retranslation of the Behavioral Examples

Sample. The retranslation judges were a different group of individuals than those who generated the critical incidents. This sample consisted of 45 NCOs and 36 officers. There were 59 males and 22 females. The average time in the Army for the sample was 8.53 years and the average amount of supervisory experience was 4.75 years. The retranslation workshops were conducted at Ft. Knox, KY.

Procedures. Retranslation provides a way of checking on the clarity of individual behavioral examples and of the dimension system. This is accomplished by asking persons familiar with the target domain to make two judgments about each behavioral example: the dimension or category to which it belongs based on its content and the effectiveness level it reflects. Examples where disagreement occurs in either category membership or rated effectiveness level may be unclear and should be revised or eliminated from further consideration. Also, confusion between two or

more categories in the sorting of several examples may reflect a poorly formed category system.

To accomplish the retranslation task, judges were provided with definitions of the 12 dimensions to aid in the sorting of behavior examples into categories and a 7-point effectiveness scale (where 1 = extremely ineffective, 4 = average, and 7 = extremely effective) to guide the effectiveness ratings. Further, the retranslation task was divided into four subtasks, each requiring a retranslation judge to evaluate 200 behavioral incidents. This division into subtasks was accomplished to keep reasonable the amount of time each judge would need to spend on the retranslation task.

Results. An initial screening of the data was undertaken to identify and delete potential random responders or individuals who obviously did not understand the retranslation task. Specifically, respondents were scored on 12 critical incidents, each of which the research staff believed were very straightforward to classify into one of the 12 dimensions. If respondents did not correctly categorize at least 50% of these incidents, they were deleted from the sample. Seven respondents out of the 81 total respondents were omitted from the sample using this criterion, leaving a total sample size of 74 for the retranslation analyses reported below.

Table 1 shows the number of behavioral examples that were reliably retranslated for each of the 12 dimensions. The acceptance points for retaining an example were greater than 50 percent for sorting the example into a single dimension, and less than 1.50 standard deviation for the effectiveness rating. These criteria left 734 of the 1,000 examples (73.4%) to be included for subsequent scale development efforts. It should be noted that the retranslation resulis indicated that all 12 of

the dimensions that resulted from the initial categorization of the incidents should be retained.

Army-Wide Scale Development

The results in Table 1 are satisfactory in that sufficient numbers of reliably retranslated examples are available to develop behavioral summary statement anchors for each dimension. Typically, a minimum of 20 reliably retranslated examples that are not highly overlapping in content is considered sufficient for defining a dimension.

We are presently in the process of developing these behavioral summary statement anchors for each Army-wide performance dimension. For each dimension, the reliably retranslated examples will be divided into three categories of effectiveness levels: low (1 - 2.49), average (2.50 - 5.49), and high (5.50 - 7). Behavioral summary statements will then be written to capture the content of the specific examples at these three effectiveness levels.

Development of the behavioral summary statements is the critical step in forming Behavioral Summary Scales. The main advantage of these scales over behaviorally anchored rating scales is that, for a particular dimension and effectiveness level, the content of all of the reliably retranslated examples is represented on the scales, not just one or two of the specific behavioral examples (Borman, 1979). Accordingly, it is more likely that a rater using the scales will be able to match observed performance with the performance descriptions that appear on the scales.

Development of Second Tour MOS-Specific Rating Scales

Second tour MOS-specific rating scales were develop- for nine jobs:

infantryman (11B), cannon crewmember (13B), tank crewmember (19E), single-channel radio operator (31C), light wheel vehicle mechanic (63B), motor transport operator (64C), administrative specialist (71L), medical specialist (91A/B), and military police (95B). The approach used for developing these rating scales differed from the approach used to develop the second tour Army-wide rating scales. Whereas the second tour Army-wide rating scales were developed using the entire sequence of behavioral summary scale procedures, development of the second tour MOS-specific rating scales involved revising the first tour MOS-specific rating scales so that they would be appropriate for evaluating second tour performance.

Behavior Analysis Workshops

Sample. A behavior analysis workshop was conducted with officers and NCOs in each of the nine target jobs to generate examples of effective, average, and ineffective second tour MOS-specific job performance. Approximately 25 individuals participated in each workshop. The participants had an average of 8.42 years in the Army and an average of 5.78 years of supervisory experience. The workshops were conducted at Ft. Knox, KY, Ft. Bragg, NC, Ft. Carson, CO, Ft. Sam Houston, TX, Ft. Gordon, GA, and Ft. Hood, TX.

Procedure. The same procedures used to generate the Army-wide behavior examples were used in the MOS-specific behavior analysis workshops. However, rather than writing examples that would be applicable to any MOS, participants were instructed to write behavior examples that were specific to the particular job for which they were writing incidents. The numbers of behavioral examples generated for each MOS were as follows: 11B (161 examples), 13B (58 examples), 19E (236 examples), 31C (212 examples), 63B (180 examples), 64C (184 examples), 71L (149 examples), 91A/B (206 examples), and 95B (234 examples).

Comparison of First Tour and Second Tour Behavior Examples

The behavior incidents were first edited to a common format. Then, they were categorized for each job using the first tour MOS-specific category system as a starting framework. If a second tour incident did not "fit" into an already existing first tour category, an entirely new category was introduced. Through this process, it was possible to determine whether the same or different categories should be used for evaluating second tour performance as were used to evaluate first tour performance. This exercise also yielded information regarding what specific category additions or deletions were necessary to comprehensively tap the second tour performance domain.

Almost all of the first tour MOS-specific rating categories were replicated in some form for the second tour jobs. For each category that was both a first tour and a second tour dimension, the next step was to examine the content of the incidents to determine whether or not the performance requirements were appreciably different for second tour than for first tour soldiers. This was an important step because although the names of the performance dimensions for first and second tour soldiers might be the same, it was at least possible that the dimension definitions or anchors might need to be modified/revised in order to make the scales appropriate for evaluating second tour performance.

For some dimensions, comparisons of the first and second tour behavior incidents indicated that more was expected of second tour soldiers performing at the average or high levels of performance than was expected of their first tour counterparts. In other cases, low level performance for a first tour soldier seemed "too low" for individuals in their second tour. Under such circumstances, the summary statement

anchors were modified to reflect the appropriate performance standards. For other dimensions, the incidents suggested that second tour soldiers were responsible for knowing how to operate and maintain more/different pieces of equipment than were the first tour soldiers. Again, this type of difference was incorporated into the second tour summary statements.

For several of the MOSs, the second tour incidents also suggested that some new, MOS-specific supervisory categories should be developed. Accordingly, preliminary summary statement anchors were written for these supervisory dimensions. In developing the categories, however, care was taken not to duplicate the Army-wide supervision categories, which would be used to evaluate individuals in all MOSs. That is, if the supervisory incidents reflected the same types of behaviors that were already being tapped by the Army-wide supervisory dimensions, then no MOS-specific supervisory dimensions were developed. Thus, the MOS-specific categories reflected aspects of supervision that were relevant only to the particular job in question. The names of the second tour supervisory performance dimensions by MOS are shown in Table 2. As it can be seen from the table, MOS-specific supervisory dimensions were developed for five of the nine MOSs.

Scale Revision Workshops

Sample. For each MOS, two scale revision workshops were conducted with 10-14 participants in each. These individuals were different from those who generated the behavior examples. Approximately half of the participants were officers and the other half were NCOs. The sample reported an average of 5.86 years in the Army and an average of 3.43

years of supervisory experience. The scale revision workshops were conducted at Ft. Bragg, NC and Ft. Carson, CO.

Procedure. The purpose of the scale revision workshops was to have subject matter experts review the proposed second tour performance categories and make any revisions to the dimension definitions and anchors that were necessary to make the scales appropriate for evaluating second tour MOS-specific performance. Participants were told that three focal questions needed to be addressed during the workshops:

- Do the dimension anchors contain material that is not relevant for evaluating second tour soldier effectiveness?
- Do the dimension anchors for various levels of effectiveness accurately reflect what would be expected of a second tour soldier performing at the ineffective, average, and effective levels of performance?
- Do the proposed dimensions tap all of the MOS-specific performance components of the second tour soldier's job?

To answer these questions, the workshop leader reviewed each dimension in detail with the workshop participants. One by one, the three summary statement anchors describing ineffective, average, and effective performance for each dimension were discussed. Participants were asked to think about second tour performance expectations and recommend any changes that they deemed necessary to make the scales maximally relevant for evaluating second tour soldiers.

Based on the input from the workshop participants, the scales were revised. In most cases, only minor wording changes were made to the summary statements. In a few cases, however, the dimensions themselves as well as their anchors were changed substantially. Substantial changes

to the dimensions were usually a result of the job requirements having actually changed since the time the first tour scales were developed and the second tour behavior incidents were collected. Workshop participants made a final review of the proposed changes to the rating scales before being dismissed.

Retranslation Workshops

Sample. For each MOS, a retranslation workshop was conducted with approximately 20 officers and NCOs. The total number of individuals participating in the retranslation workshops across all MOS was 193. Workshop participants were again different from those who generated the critical incidents and those who reviewed and revised the proposed second tour rating scales. For this sample, the average time in the Army was 7.34 years and the average amount of supervisory experience was 3.96 years. Retranslation workshops were conducted at Ft. Carson, CO and Ft. Lewis, WA.

Procedure. The purpose of the retranslation workshops was to check on the intended effectiveness levels of the behavioral summary statements anchoring each MOS-specific performance dimension as well as to check on the dimension structures themselves. It is important to clarify that rather than retranslating individual behavior examples (as was the case with the Army-wide retranslation workshops described above), participants were asked to retranslate the actual summary statements that would be used to anchor the rating scale dimensions.

Recall that there were three summary statements anchoring each dimension: one describing low level or ineffective performance, one describing middle level or average performance, and one describing high level or effective performance. Participants were provided with definitions of each dimension and a booklet containing the summary

statements listed in a random order. They were asked to make two judgments about each summary statement: the dimension or category to which it belonged based on its content and the effectiveness level it represented from 1 (very ineffective) to 7 (very effective). The number of dimensions for the different MOS ranged from a minimum of seven (for the 71L's) to a maximum of 14 (for the 95B's). Thus, judges were required to make from 21 to 42 judgments for this retranslation task.

Results. Again, an initial screening of the data was undertaken to identify and delete potential random responders or individuals who obviously did not understand the retranslation task. For each MOS, respondents were scored on approximately 10 critical incidents each of which the research staff believed were very straightforward to classify into one of the performance dimensions. If respondents did not correctly recategorize at least 50% of their incidents, they were deleted from the sample. Of the 193 total participants in the retranslation workshops, 22 were excluded from the retranslation analyses reported below.

For almost all (98%) of the summary statements for all of the nine MOSs, at least half of the retranslation sample placed them in the intended category, and for 92% of the statements, more than 75 percent of the sample categorized them as intended. The mean effectiveness level was also very close to the intended effectiveness level for most of the summary statements. That is, if the statement was intended to be a low level or ineffective anchor, its mean effectiveness level was about a 1. For those intended to be a middle level or average anchor, the mean effectiveness level was about a 4, and for those intended to be a high level or effective anchor, the mean effectiveness level was about 7.

There were a few statements (about 14% across all MOS), however, for which there was some discrepancy between the mean effectiveness level and the intended effectiveness level (i.e., the effectiveness rating was more than one point away from the intended effectiveness level). Revisions were made to such statements to ensure that they reflected the proper effectiveness levels.

Discussion

The MOS specific second tour rating scales appear ready for field testing. Retranslation results indicate that the category system for each MOS's scales and the effectiveness levels reflected in the summary statements anchoring the rating categories will provide a comparatively unambiguous rating format for evaluating second tour soldier performance in these MOS. The Army-wide scale development effort is nearing completion. All that remains is preparation of behavioral summary statements to anchor the three general levels of effectiyeness for each of the Army-wide dimensions. The rest of this discussion focuses on the "shortcut" method used here to develop second tour MOS-specific scales and inferences that might be made about the nature of the second tour soldier job based on the content of the behavioral incidents gathered.

Comments on the "Shortcut" Method for MOS-Specific Scale Development

One lesson learned from the MOS-specific scale development effort is that a procedure less time consuming than the usual behavioral scale development sequence may be very effective when behavior-based rating scales for a similar job are already available. The typical approach for constructing such scales is to elicit large numbers of behavioral examples, develop dimensions based on the content of the examples, have the examples retranslated into those dimensions and according to effectiveness level, and write behavioral summary statements for each

performance level on each dimension. In addition, the summary statements are often reviewed by job experts before the scales are put in final form.

Because, the first tour behavioral rating scales were available for each of the nine MOS and because the second tour performance requirements were reported to be similar in many ways to first tour requirements, it seemed appropriate in our research to simplify the MOS-specific scale development procedures. Accordingly, and as mentioned previously in this paper, the first tour scales were used as a starting point in the present effort. Those parts of the scales requiring changes were revised utilizing a relatively small number of performance examples and a group of job experts working directly on the scales' summary statements. This shortened procedure reduced considerably the time and expense needed for rating scale development without reducing the quality of the scales, as was apparent from the favorable retranslation results for the final summary statements.

On the Nature of the Second Tour NCO Job: Technical and Supervisory Duties

An important job content-related issue addressed by these rating scale development results concerns the nature of the second tour NCO job. Specifically, second tour soldiers have a variety of performance requirements, some involving technical aspects of the job and others relating to supervisory duties. Second tour NCOs both perform and supervise the work. Data gathered in the present effort provide some information relevant to determining the salience of the technical versus supervisory elements of these jobs.

In particular, the content of the performance examples or incidents gathered for the nine MOSs should reveal estimates of the relative importance of the technical and supervisory aspects of the second tour soldier job. Consideration of how these performance incidents were elicited will clarify why this is so. Recall that the NCOs and their supervisors from each of the target MOSs were asked in a workshop setting to record behavioral incidents they recalled from observing second tour soldiers working in these MOSs. Workshop participants were told that the incidents could refer to any part of the job for that MOS, so we would expect the content of a large number of incidents gathered inductively in this manner should representatively sample the different elements of the job.

More precisely, we would expect that the performance incidents elicited this way would reflect a representative sample of the job content related to performance requirements, what it takes to be effective on these jobs (rather than, for example, the time spent on different job activities). This is because the behavioral analysis method draws out incidents whose content relates to effectiveness on the job. As mentioned previously in this section, we did not gather a large number of performance incidents for each MOS, but the incidents we did collect, across all MOSs, and the Army-wide incidents appear to yield a sufficient sampling to provide a look at the issue of job content, technical versus supervisory, related to performance requirements on these jobs.

Table 3 shows the percent supervisory performance incidents as judged by our research staff, for each of the nine MOSs, along with the

total percentage of MOS-specific incidents that were supervisory in nature, across all nine MOSs. Referring to individual MOSs, second tour infantrymen and light wheel vehicle mechanics seem to do the most supervising, while tank crewman and vehicle operators are involved least in supervising soldiers.

Comparing Table 3 with Table 2 notice that our decision to develop (or not to develop) MOS-specific supervisory categories for each MOS was not directly related to the percentage of supervisory incidents gathered for that MOS. Rather, as mentioned previously, MOS-specific supervisory categories were developed only when the incidents for that MOS reflected aspects of supervision which were not tapped by the Army-wide supervisory dimensions.

Table 4 presents a more detailed analyses of the MOS-specific supervisory performance incidents. Also shown in Table 4 is the percentage of the 734 Army-wide incidents reliably retranslated into the supervisory performance dimensions in the Army-wide scale development effort. Although the total percentages of MOS-specific and Army-wide supervisory incidents are reasonably close (27.1% and 30.5%), the distribution of these incidents to individual supervisory categories is very uneven across the two sources of incidents. The vast majority of the MOS-specific supervisory incidents fall in the Organizing, Supervising, Monitoring, and Correcting dimension, whereas supervisory incidents are more uniformly spread across all three categories in the Army-wide case. The reason for so few MOS-specific incidents in the "Showing Concern" dimension is probably due to the more generic nature of

that dimension and the instructions to MOS-specific workshop participants to focus on performance examples relevant only to the target MOS. It is not clear why there are differences in the patterns of incidents for the other two supervisory categories, although sampling error is certainly a possible reason for such differences.

At any rate, results in Tables 3 and 4 suggest that indeed the second tour soldier job has performance requirements in both the technical and supervisory areas. For most of the MOSs, roughly one-quarter to one-third of the performance demands are likely to be supervisory in nature, with the rest in the technical arena. This finding has, of course, important implications for soldier selection, as well as for the training and retention of second tour Army personnel. Selection concerns need to focus on personal characteristics relevant to supervisory success in addition to aptitudes and abilities important for obtaining technical knowledge and skills necessary for the technical aspects of the job. Training must emphasize skill-building instruction and on-the-job experiences related to technical and supervisory aspects of the job. And, retention of second tour soldiers with skills and potential in both areas should be explicitly encouraged. Project A researchers are attending to these implications in continuing efforts to improve the overall effectiveness of the U.S. Army.

References

Borman, W. C. (1979) Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 412-421.

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1987). Development of a model of soldier effectiveness. ARI Technical Report 741. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986). Development and field test of Project A task-based MOS-specific criterion measures. ARI Technical Report 717. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

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

Davis, R. H., Davis, G., & Joyner, J., and deVera, M. V. (1986). Development and field test of job relevant knowledge tests for selected MOS. ARI Technical Report 757, in press. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Peterson, N. G. (Editor). (1986). Development and field test of the trial battery for Project A. ARI Technical Report 739. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Pulakos, E. D., & Borman, W. C. (Editors). (1986). Development and field test of the Army-wide rating scales and the rater orientation and training program. ARI Technical Report 716. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. (1986). Development and field test of behavioral anchored rating scales for nine MOS. ARI Technical Report 776. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Table 1. Summary of Reliably Retranslated Second Tour Army-Wide Incidents by Category

Category	#/% of Incidents
A. Displaying Technical Knowledge/Skill	55/7%
B. Displaying Effort, Conscientiousness, and Responsibility	168/23%
C. Organizing, Supervising, Monitoring, and Correcting Subordinates	99/13%
D. Training and Developing	63/9%
E. Showing Consideration and Concern for Subordinates	62/8%
F. Following Regulations/Orders and Displaying Proper Respect for Authority	59/8%
G. Maintaining Own Equipment	21/3%
H. Displaying Honesty and Integrity	59/8%
I. Maintaining Proper Physical Fitness	32/4%
J. Developing Own Job/Soldiering Skills	43/6%
K. Maintaining Proper Military Appearance	27/4%
L. Controlling Own Behavior Related to Personal Finances, Drugs/Alcohol, and Aggressive Acts	46/6%

Table 2. Supervisory Performance Categories for Second Tour MOS-Specific Scales

MOS	Performance Category Name
11B	Supervising Soldiers in the Field
	Leading the Team
13B	None
19E	Assuming Supervisory Responsibilities in Absence of Tank Commander
31C	Managing the RATT Rig
63B	Checking Repairs Made by Other Mechanics
64C	None
71L	None
91A/B	None
95B	Leading the Team in a Tactical Environment

Table 3. Percent Supervisory Performance Incidents From MOS-Specific Workshops

MOS	Total Number of Incidents	Number of Supervisory Incidents	Percent Supervisory MOS-Specific Incidents
11B	159	71	44.7%
13B	57	13	22.8%
19E	236	27	11.4%
31C	212	49	23.1%
63B	180	76	42.2%
64C	184	31	16.8%
71L	156	36	23.1%
91A	89	33	37.1%
95B	234	73	31.2%
Totals	1507	409	27.1%

Table 4. Numbers and Percent Performance Incidents By Supervisory Category

	MOS-Specific Incidents		Army-Wide Incidents	
	N	%	N	%
1. Organizing, Supervising, Monitoring, and Correcting Subordinates	310	20.6%	99	13.5%
2. Training and Developing Subordinates	82	5.4%	63	8.6%
3. Showing Consideration and Concern for Subordinates	17	1.1%	62	8.4%
Totals	409	27.1%	224	30.5%

GETTING ANSWERS TO THE RIGHT QUESTIONS:

JOB ANALYSIS STRATEGY

Michael G. Rumsey

U.S. Army Research Institute

Presented on Symposium,

"Junior Noncommissioned Officer Job Requirements;
Where Does Leadership Fit In?"

At the Annual Convention of the
American Psychological Association
New York

August 1987

Getting Answers to the Right Questions:

Job Analysis Strategy

To many, the words "job analysis" fail to generate a sudden surge of excitement. Instead, they may evoke images of mindless automatons poring over endless lists of task statements. The practitioners in this field should be forgiven if they sometimes find themselves identifying with Rodney Dangerfield.

Such an image of job analysis in fact poorly represents the nature of the activity. It encompasses issues which are challenging, stimulating and critically important. Consider this situation. You are building an enlisted selection and classification system for the entire U.S. Army. You want to test the validity of this system in as rigorous a manner as possible, so you have set about to build a comprehensive set of criterion measures to capture soldier performance in both the first and second tours of duty. Your mission is partially complete; you have finished development of first tour measures.

Now, as you approach development of second tour measures, you realize answers to several key questions are needed before you can proceed. First, what should be the content of these measures? Second, are separate measures needed for each job? Or are the jobs so similar that the same measures can be applied to all? Third, to what extent can first tour measures be used in second tour? You do not want to squander valuable resources to develop new second tour measures if there is really no major difference between first and second tour performance. Fourth, what kinds of measurement methods are needed? These need to be suitable to the job requirements.

Clearly, you now need job analysis information. The challenge here is to develop that job analysis strategy which will not only identify and prioritize job components, but which will furthermore provide sufficient information to ensure a maximally effective set of performance measures. Such a strategy should yield as comprehensive a job picture as possible. Multiple methods are to be preferred as likely yielding more complete information than might a single method. To the degree feasible, all relevant and useful sources of information should be consulted.

As you have probably guessed, the scenario I have been describing is not merely a hypothetical one. It is essentially the situation we found ourselves in as we prepared to analyze nine second tour jobs in Project A, a large scale project to develop performance-based selection and classification measures for the Army. Before the other members of this panel tell you what we have been learning from these analyses, I would like to spend the next few minutes describing the overall strategy that guided our efforts.

At the outset of this project, we had advanced a general strategy for job analysis designed to provide both good overall job coverage and a basis for discriminating between good and poor performers. A multimethodological approach was adopted which incorporated two of the three basic types of job analysis methods identified by Ash (1982): task-based and behavior-based. The task-based approach involved heavy reliance on existing job information, supplemented by interviews with cognizant subject matter experts, to first identify a consolidated domain of all tasks within a job. From this domain, a smaller set of tasks was to be identified which could best represent the full

domain for testing purposes. Finally, the tasks in the smaller set were divided into discrete steps (Campbell, Campbell, Rumsey, & Edwards, 1985).

The behavior-based approach involved workshops in which subject matter experts on the job generated examples of good, poor and average performance. These examples were then clustered into dimensions. (Toquam, McHenry, Corpe, Rose, Lammlein, Kemery, Borman, Mendel, & Bosshardt, in preparation).

This general approach was, in our judgment, reasonably successful for the analysis of first tour jobs. It led to the measures which were judged by responsible Army proponents were to provide adequate job coverage and which provided reasonable discrimination among those tested. But the job requirements at the first tour level were relatively uncomplicated. A soldier was essentially expected to be able and willing to do the work required. Among the second tour soldiers we would be examining, many would have advanced to a junior non-commissioned officer level. The available literature (Hebein, Kaplan, Miller, Olmstead & Sharon, 1984; Wallis, Korotkin, Yarkin-Levin, Schemmer, & Mumford, 1985), as well as preliminary soldier interviews, indicated that at this level soldiers would have supervisory as well as technical job requirements. Would the first tour job analysis approach still suffice for soldiers required to assume responsibility for the work and behavior of others?

It is as true in job analysis as elsewhere that the answers one gets is to no small degree a function of the questions one asks. In our behavior-based approach we had been asking essentially two kinds of questions: what are critical behaviors for effective performance on a specific job and what

are critical behaviors for effective performance on any type of Army job? These questions seemed sufficiently encompassing to capture both supervisory and non-supervisory job requirements.

Our real concern was what we would find, or fail to find, using the task-based approach. Let us for the moment split second tour requirements into two categories--technical and supervisory; recognizing that such a dichotomy represents a gross oversimplification. In other contexts, we will be using the word "technical" in a much more restricted way.

A task is, by one definition, an observable, measurable action, with a definite beginning and end, which is performed for a relatively short period of time (DeVries, Eschenbrenner & Ruck, 1980, pp. 10, 13). This definition fits technical tasks reasonably well; in fact, the task-based approach seems principally designed to generate tasks which are technical in nature. It was our expectation that this approach would provide satisfactory coverage of the technical domain.

We had no such expectation with respect to the supervisory domain. Supervisory behaviors tend to be continuous rather than discrete, are not easily observable and measurable, and are difficult to fix in time. Since it is difficult to translate leader behaviors into tasks, those generating task inventories may omit such behaviors entirely or represent them inadequately.

We felt the task-based approach provided useful information and should be included in our overall strategy. The dilemma was how to insure that the task lists generated provided adequate representation of supervisory job requirements.

Our basic strategy was simply to expand the sources we used to generate our consolidated task list. Fortunately, sources were available which, when combined, gave us reasonable confidence that we were covering the supervisory domain. Alma Steinberg, a contributor to the next paper, and her colleagues (Steinberg, van Rijn & Hunter, 1986) had, through extensive interviews, generated a comprehensive task list focused on leader requirements from the junior NCO to the senior officer level. Ilene Gast, our next speaker, had generated a list of leader tasks based on critical incidents which was less exhaustive than the list generated by Steinberg but which tended to be more focused at the junior NCO level.

Following a preliminary data collection effort which provided more information about the relevance of tasks on both lists for second tour soldiers, the two lists were merged into one through a process designed to retain the most desirable characteristics of each.

At this point, we had a strategy which we believed could provide the information we sought about measurement content and method, the extent to which measures could be collapsed across jobs, and the extent to which first tour measures were appropriate for second tour soldiers. The following papers will explore how this strategy was applied and what answers were generated by it.

References

Ash, R. A. (1982). Job elements for task clusters: Arguments for using multi-methodological approaches and a demonstration of their utility. Public Personnel Management Journal, 11, 80-90.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1985). Development and field test of task-based MOS-specific criterion measures. (Tech. Rep. No. 717). Alexandria, VA: U.S. Army Research Institute.

DeVries, P. B., Eschenbrenner, A. J., & Ruck, H. W. (1980). Task analysis handbook (Tech. Rep. No. 79-45). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Fine, S. A. (1974). Functional job analysis: An approach to a technology for manpower planning. Personnel Journal, 53, 813-818.

Hebein, J., Kaplan, A., Miller, R., Olmstead, J., & Sharon, B. (1984). NCO leadership: Tasks, skills, and functions. (Research Note No. 84-95). Alexandria, VA: Human Resources Research Organization.

Steinberg, A. G., van Rijn, P., & Hunter, F. T. (1986). Leader requirements task analysis. Paper presented at the 28th Annual Military Testing Association Conference, Mystic, Connecticut.

Toquam, J. L., McHenry, J. J., Corpe, V. A. Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (in preparation). Development and field test of behaviorally anchored rating scales for nine MOS.

Wallis, M. R., Korotkin, A. L., Yarkin-Levin, K., Schemmer, F. M., & Mumford, M. D. (1986). Leadership job dimensions and competency requirements for commissioned and noncommissioned officers: Remediation of inadequacies in existing data bases. (Research Note No. 86-20) Alexandria, VA: U.S. Army Research Institute.

930203