DTIC FILE COPY

ARI Research Note 87-54

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

## ANNUAL REPORT, 1985 FISCAL YEAR- SUPPLEMENT TO ARI TECHNICAL REPORT TR 746

John P. Campbell, Editor
Human Resources Research Organization

American Institutes for Research
Personnel Decisions Research Institute
Army Research Institute

for

Selection and Classification Technical Area
Lawrence M. Hanser, Chief

MANPOWER AND PERSONNEL RESEARCH LABORATORY
Newell K. Eaton, Director

AD-A188 267

DTIC
ELECTE
DEC 0 2 1987
S D
H

U. S. Army

Research Institute for the Behavioral and Social Sciences

October 1987

Approved for public release; distribution unlimited.

87 11

# U. S. ARMY RESEARCH INSTITUTE

# FOR THE BEHAVIORAL AND SOCIAL SCIENCES

## A Field Operating Agency under the Jurisdiction of the

## Deputy Chief of Staff for Personnel


EDGAR M. JOHNSON
*Technical Director*

WM. DARRYL HENDERSON
COL, IN
Commanding



Research accomplished under contract
to the Department of the Army

Human Resources Research Organization

Technical review by

Michael G. Rumsey
Clinton B. Walker



| Accession For | |
|---|---|
| NTIS GRA&I | ☒ |
| DTIC TAB | ☐ |
| Unannounced | ☐ |
| Justification | |
| By | |
| Distribution/ | |
| Availability Codes | |
| Dist | Avail and/or Special |
| A-1 | |

This report, as submitted by the contractor, has been cleared for release to Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or other reference services such as the National Technical Information Service (NTIS). The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

| | |
|---|---|
| 1. REPORT NUMBER: ARI Research Note 87-54 | 2. GOVT ACCESSION NO.: AD-A188267 |
| 3. RECIPIENT'S CATALOG NUMBER | |
| 4. TITLE (and Subtitle): IMPROVING THE SELECTION, CALSSIFICATION AND UTILIZATION OF ARMY ENLISTED PERSONNEL: Annual Report, 1985 Fiscal Year - Supplement to ARI Technical Report 746 | 5. TYPE OF REPORT & PERIOD COVERED: Final Report October 84 - September 85 |
| | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s): John P. Campbell (editor) | 8. CONTRACT OR GRANT NUMBER(s): MDA 903-82-C-0531 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: Human Resources Research Organization, 1100 South Washington Street, Alexandria, VA 22314-4499 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 20263731A792, 2.3.2. C1 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS: U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333-5600 | 12. REPORT DATE: October 1987 |
| | 13. NUMBER OF PAGES: 502 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): - - | 15. SECURITY CLASS. (of this report): Unclassified |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: n/a |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited.

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

- -

18. SUPPLEMENTARY NOTES

Lawrence M. Hanser, contracting officer's representative and technical point of contact. Other organizations assisting in preparation of this report were the American Institutes for Research and the Personnel Decisions Research Institute.

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

| | |
|---|---|
| Army-Wide Measures | Predictor Measures |
| Classification | Project A |
| Criterion Measures | Ratings |
| Hands-On Tests | Selection |
| Knowledge Tests | Soldier Effectiveness |

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

The materials presented in this research note were prepared as part of "Project A", the Army's current large-scale manpower and personnel effort, which seeks to improve the selection, classification, and utilization of Army Enlisted personnel. This note supplements ARI Technical Report 746, the Project Annual report for the 1985 Fiscal Year. It augments that report by providing copies of a set of technical papers that were prepared during the year reporting in detail on phases of the project research method, and the results.

## EDITORS' PREFACE

In the course of executing the mainline research program of Project A, it has always been an accepted--indeed priority--practice to find mechanisms and means for communicating and sharing early and/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 reports in this Supplement to the Fiscal Year 1985 Annual Report are presented in chronological order. Most of them 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 which can reveal the process and flow of the overall research program being accomplished collegially by the U.S. Army Research Institute 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

## EDITORS' PREFACE

In the course of executing the mainline research program of Project A, it has always been an accepted--indeed priority--practice to find mechanisms and means for communicating and sharing early and/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 reports in this Supplement to the Fiscal Year 1985 Annual Report are presented in chronological order. Most of them 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 which can reveal the process and flow of the overall research program being accomplished collegially by the U.S. Army Research Institute 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, 1985 FISCAL YEAR
## SUPPLEMENT TO ARI TECHNICAL REPORT 746

## 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 746, the Project Annual Report for the 1985 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 which the U.S. Army has undertaken to develop an improved personnel selection and classification system for 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 which will validly predict carefully developed measures of job performance. The project addresses the 675,000-person enlisted personnel system of the Army, encompassing several hundred different military occupations.

This research program began in 1980, when the U.S. Army Research Institute (ARI) started planning the extensive research effort that would be needed to develop the desired system. In 1982 a consortium led by the Human Resources Research Organization (HumRRO) and including the American Institutes for Research (AIR) and the Personnel Decisions Research Institute (PDRI) was selected by ARI to undertake the 9-year project. The total project utilizes the services of 40 to 50 ARI and consortium researchers working collegially in a variety of specialties, such as industrial and organizational psychology, operations research, management science, and computer science.

The specific objectives of Project A are to:

- Validate existing selection measures against both existing and project-developed criteria. The latter are to include both Army-wide job performance measures based on newly developed rating scales, and direct hands-on measures of MOS-specific task performance.
- Develop and validate new selection and classification measures.
- Validate intermediate criteria (e.g., performance in training) as predictors of later criteria (e.g., job performance ratings), so that better informed reassignment and promotion decisions 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 operational selection and classification decisions.

The research design for the project 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 Army accessions in fiscal years (FY) 1981 and 1982 were evaluated to explore the relationships between the scores of applicants on the Armed Services Vocational Aptitude Battery (ASVAB), and their subsequent performance in training and their scores on the first-tour Skills Qualification Tests (SQT).

In the second iteration, a concurrent validation design will be executed with FY83/84 accessions. As part of the preparation for the Concurrent Validation, a "preliminary battery" of perceptual, spatial, temperament/personality, interest, and biodata predictor measures was assembled and used to test several thousand soldiers as they entered in four Military Occupational Specialties (MOS). The data from this "preliminary battery sample" along with information from a large-scale literature review and a set of structured, expert judgments were then used to identify "best bet" measures. These "best bet" measures were developed, pilot tested, and refined. The refined test battery was then field tested to assess reliabilities, "fakability," practice effects, and so forth. The resulting predictor battery, now called the "Trial Battery," which includes computer-administered perceptual and psychomotor measures, will be administered together with a comprehensive set of job performance indices based on job knowledge tests, hands-on job samples, and performance rating measures in the Concurrent Validation.

In the third iteration (the Longitudinal Validation), all of the measures, refined on the basis of experience in field testing and the Concurrent Validation, will be administered in a true predictive validity design. About 50,000 soldiers across 20 MOS will be included in the FY86-87 "Experimental Predictor Battery" administration and subsequent first-tour measurement. About 3500 of these soldiers are estimated for availability for second-tour performance measurement in FY91.

For both the concurrent and longitudinal validations, the sample of MOS was specially 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. Sample sizes are sufficient so that race and sex fairness can be empirically evaluated in most MOS.

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

The development and revision of the wide variety of predictor and criterion measures reached the stage of extensive field testing during FY84 and the first half of FY85. These field tests resulted in the formulation of the test batteries to be used in the comprehensive Concurrent Validation program which was initiated in FY85. Various reports on specific aspects of the field tests have been issued.

Activities and progress during the first two years of the project were reported for FY83 in ARI Research Report 1347 and its Technical Appendix, ARI Research Note 83-37, and for FY84 in ARI Research Report 1393 and its related reports, ARI Technical Report 660 and ARI Research Note 85-14. Other publications on specific activities during those years are listed in those annual reports. The annual report on project-wide activities during FY85 is presented in ARI Technical Report 746. The technical papers reproduced in this Research Note serve as documentation for various FY85 activities.

## CONTENTS


| | Page |
|---|---|
| Validation of the Army's Military Applicant Profile (MAP) Against an Expanded Criterion Space. . . . . Clinton B. Walker (ARI) | 1 |
| Influence of Soldiers' Experiences With Supervisors on Performance During the First Tour. . . . . Leonard A. White, Ilene F. Gast, Helen M. Sperling, and Michael G. Rumsey (ARI) | 11 |
| The Fakability of the Army's Military Applicant Profile (MAP) . . . . Clinton B. Walker (ARI) | 17 |
| Assessing the Utility of a Personnel/Classification System . . . . . Robert Sadacca and John P. Campbell (HumRRO) | 27 |
| Performance Ratings as Criteria: What Is Being Measured? . . . . . Walter C. Borman (PDRI), Leonard A. White and Ilene F. Gast (ARI), and Elaine D. Pulakos (PDRI) | 79 |
| Criterion Reduction and Combination via a Participative Decision-Making Panel. . . . . John P. Campbell and James H. Harris (HumRRO) | 103 |
| Measurement of Entry-Level Job Performance . . . . . Newell K. Eaton (ARI) | 159 |
| Problems, Issues, and Results in the Development of Temperament, Biographical, and Interest Measures. . . . . Leaetta M. Hough, Bruce N. Barge, Janis S. Houston, Matt K. McGue, and John D. Kamp (PDRI) | 183 |
| Problems, Issues, and Results in the Development of Computerized Psychomotor Measures . . . . . Jeffrey J. McHenry and Matthew K. McGue (PDRI) | 233 |
| Measurement of Test Battery Value for Selection and Classification . . Donald H. McLaughlin (AIR) | 261 |
| Overall Strategy and Methods for Expanding the Measured Predictor Scale. . . . . Norman G. Peterson (PDRI) | 283 |
| Advantages and Problems With Using Portable Computers for Personnel Measurement. . . . . Rodney L. Rosse and Norman Peterson (PDRI) | 305 |
| Modeling the Selection Process to Adjust for Restriction in Range. . . Paul G. Rossmeissl (ARI) and David A. Brandt (AIR) | 325 |

**CONTENTS (Continued)**


Page

Comparing Work Sample and Job Knowledge Measures . . . . . 337
Michael G. Rumsey (ARI), William C. Osborn and Patrick Ford (HumRRO)

Development of Cognitive/Perceptual Measures: Supplementing the ASVAB. . . . . 377
Jody L. Toquam, Marvin D. Dunnette, VyVy Corpe, Jeffrey J. McHenry, Margaret A. Keyes, Matthew K. McGue, Janis S. Houston, Teresa L. Russell, and Mary Ann Hansen (PDRI)

Expanding the Measurement of Predictor Space for Military Enlisted Jobs. . . . . 435
Hilda Wing (ARI)

Development of an Index of Maximum Validity Increment for New Predictor Measures . . . . . 439
Lauress L. Wise (AIR) and Karen J. Mitchell (ARI)

Personal Constructs, Performance Schemata, and "Folk Theories" of Subordinate Effectiveness: Explorations in an Army Officer Sample . . 461
Walter C. Borman (PDRI)

# VALIDATION OF THE ARMY'S MILITARY APPLICANT PROFILE (MAP) AGAINST AN EXPANDED CRITERION SPACE

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

November 1984

Presented at the Military Testing Association in
Munich, Germany

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

# VALIDATION OF THE ARMY'S MILITARY APPLICANT PROFILE (MAP) AGAINST AN EXPANDED CRITERION SPACE [1]

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

This research tests the predictive validity of the U.S. Army's Military Applicant Profile (MAP) against a more varied criterion space than in the past. MAP is a multiple-choice background questionnaire which is now used to screen male volunteers who have not completed high school (NHSG). Having been validated against a criterion of successful completion of the first six months of service, versus discharge for failures to adapt (i.e., adverse causes like drug use or court martial), MAP is meant to minimize such failures. Its 53 scored items form one scale of 0 to 106 points on which scores of 62 or higher pass.

In 1977, the current scoring key was empirically derived from the correlations of item-level responses with the criterion of six-month success. Respondents then were 2,280 male NHSG trainees in their first three days of service at two Army posts. Data for the present research come from over 8,000 applicants who took the MAP in Fiscal Years (FY) 1981 and 1982 as a pre-induction screen. In that same period, separate research was being done on 9,600 trainees to develop new forms of MAP for use with females and high-school graduates (Erwin, 1984).

In past research (Eaton, Weltin, & Wing, 1982), a strong direct relation has been observed between six-month success and new recruits' scores on MAP. That finding and a similar one from the Navy's biodata screen (Atwater & Abrahams, 1983) suggest that adaptability is a continuous variable. Thus, criterion measures of successful adaptation should vary directly with individuals' scores on such measures. Owing to the likelihood of differences between the development sample for MAP and the present sample, correlations somewhat lower than the .3-.4 range of the original research (Frank & Erwin, 1978) were expected here.

The present work introduces the criteria of success beyond six months, tenure, and promotions. Test scores in training and on the job are included for exploratory purposes, but such cognitive outcomes should not correlate highly with our biodata predictor. In two other respects this work breaks new ground. First, it examines the effectiveness of the scoring key in a period of three to five years after its development. Second, MAP's validity is tested on applicants for the first time.

---

All opinions expressed in this paper are those of the author and do not necessarily reflect the official positions or policies of the U. S. Army Research Institute or the Department of the Army.

[1]Thanks go to Drs. Karen Mitchell and Paul van Rijn for their helpful reviews of a draft of this paper and to Winnie Young for creating the dataset.

## METHOD

### Cases

The cases were volunteers for the enlisted Army who entered in FY 81 and 82 after taking the MAP at any of 69 Military Enlistment Processing Stations (MEPS). During that period, 17-year old male NHSG were the official target of this pre-induction screen. In all, 10,415 machine-scorable MAP answer sheets were scanned to provide the predictor scores.

The cases consisted of all applicants for enlistment in the Regular Army between October, 1980, and September, 1982, who took the MAP, whose MAP answer sheets were in condition to be machine scored, and for whom matching records were found in the Army Applicant/Accession Files. About 500 answer sheets were not in scorable condition and 2,177 cases could not be matched. We have no reason to think that those missing cases were atypical. Cases with regular high school diplomas were excluded from all computations, but NHSG who were older than 17 (n=195) were not. Females (n=266) were excluded because MAP had been keyed only on males. For correlations with criteria, the 1,763 cases that did not enter the Army were excluded, as were 334 early discharges for benign causes, such as hardship. After these exclusions, the sample for analysis had 5,941 cases.

### Data

Data were available from Army Applicant/Accession Files, a special file of training data (2,156 cases), the Army Enlisted Master Files for FY 81/82, and DMDC gain/loss records. Information on dates and types of discharges was available through September, 1983. Variables were chosen for analysis based on their freedom from missing data, lack of extreme criterion splits, and credibility under cross-checking.

Criterion variables were tenure, success of service, rank, scores on training tests, and scores on Skills Qualification Tests (SQT). Tenure was simply the number of days from date of entry to date of separation. For cases who were still on active duty as of our latest information (viz., 9/30/83), that date was used as their date of separation. Success was defined as "absence of an adverse discharge" (i.e., an early separation for bad cause). Success was examined at two points: the end of the first six months of service, which is the end of entry-level training, and the end of the dataset. The latter criterion is called one- to three-year success, because of the boundaries in this dataset on possible successful tenures. A case which received an adverse discharge during the seventh through twelfth months of service was in the positive criterion group on the former measure and the negative one on the latter. Rank was treated as an interval-level variable with values of 1 through 5, for E1 through E5. Training and SQT scores had been standardized to a scale of 0 to 100.

### Analyses

Validities were computed in terms of Pearson and point-biserial correlations. For the criteria of tenure and rank, which were limited by time in service, the date of entry (i.e., a measure of the opportunity to build tenure and rank) was partialled out in computing validities.

As a check on the sensitivity of MAP to extremes in adaptability, mean MAP scores were examined for the 402 accessions who had received preinduction moral waivers and for the 649 cases who had had absences without leave (AWOL). Also, the moderating effect of MOS was tested by comparing the validities for the 14 MOS which had more than 100 cases in the dataset. In these MOS, the numbers ranged from 101 to 838, the median being 196.

## RESULTS

For the 7,820 non-graduate male applicants, the means for total MAP scores and AFQT percentiles were 71.45 (SD=7.27) and 50.91 (SD=15.80), respectively. Ninety-two percent of the sample made a passing score on MAP. Of those who did not enter the Army, 77.5% had passed MAP, while 4.4% of those who entered had failed it. Among the 5,941 accessions, MAP totals and AFQT percentiles averaged 72.14 (SD=6.55) and 51.15 (SD=15.34), while the non-entrants averaged 68.84 (SD=9.13) and 49.93 (SD=17.54). The Pearson correlation of MAP scores and AFQT percentiles was +.05.

The mean scores on MAP for the 860 who were adversely discharged within the first six months was 71.37 (SD=6.46), while the successful trainees averaged 72.31 (SD=6.53). For the criterion of one- to three-year success, the MAP scores were essentially the same as these.

After six months of service 85.5% of the accessions had been successful and 14.5% had received adverse discharges. As for one- to three-year success, the split was 57% vs. 43%. Thus a full two-thirds of the adverse separations happened after training, as shown in the average tenure of adverse discharges: 363 days. In contrast, successful soldiers averaged 764 days of service.

First order Pearson correlations of total MAP score with tenure, rank, SQT, and training tests ranged from .01 to .07. When the effect of entry date was removed, the correlations of MAP with tenure and rank were, in order, .068 and .043. Point biserials of MAP scores with the dichotomous criteria of six-month and one- to three-year success were .05 and .07.

Statistical adjustment of validities for the restriction of range in MAP had little effect since the unrestricted standard deviation was only one-tenth greater than that for the accessions. If the cut score had been 67 in these data, vice 62, the validities (again corrected for range restriction) against the same criteria would have been as high as .13 at the cost of excluding 18% of the accessions. The pattern of data underlying these correlations is shown in Table 1, where means and standard deviations on the criteria are given for six intervals of score on MAP. In Figure 1, the non-linearity of the relation between MAP scores and criteria is seen, outcomes being higher than expected for failing MAP scores.

Next, data for cases with preinduction moral waivers, AWOLs, and adverse discharges were examined separately. The 402 waiver cases scored less than one point lower (.11 standard deviations) on MAP than all other accessions (71.47 vs. 72.19), but they had a rather lower rate of success

Table 1.

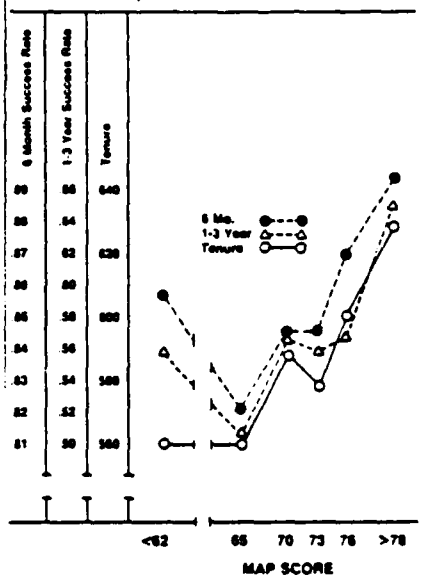
Means, Standard Deviations, and n's for AFQT and Criteria by Level of Score on the Military Applicant Profile

| Criterion | | MAP Score | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 0-61 | 62-67 | 68-71 | 72-74 | 75-78 | 79+ | Total |
| AFQT | $\bar{x}$ | 21 | 35 | 55 | 65 | 82 | 93 | 51 |
| Percentile | SD | 14 | 15 | 16 | 15 | 10 | 10 | 18 |
| | n | 356 | 1,064 | 1,266 | 1,166 | 1,264 | 706 | 5,866 |
| Six-month | $\bar{x}$ | .26 | .33 | .36 | .39 | .47 | .59 | .39 |
| Success | SD | .44 | .48 | .48 | .49 | .50 | .49 | .49 |
| (0 or 1) | n | 282 | 1,066 | 1,266 | 1,166 | 1,266 | 706 | 5,866 |
| 1-3 year | $\bar{x}$ | .36 | .51 | .57 | .66 | .77 | .86 | .57 |
| Success | SD | .50 | .50 | .50 | .48 | .42 | .35 | .49 |
| (0 or 1) | n | 282 | 1,066 | 1,266 | 1,166 | 1,266 | 706 | 5,866 |
| Tenure | $\bar{x}$ | 583 | 683 | 691 | 699 | 699 | 699 | 691 |
| (days) | SD | 399 | 399 | 399 | 391 | 399 | 370 | 391 |
| | n | 282 | 1,066 | 1,266 | 1,166 | 1,266 | 706 | 5,866 |
| Rank | $\bar{x}$ | 2.1 | 2.1 | 2.2 | 2.1 | 2.1 | 2.2 | 2.1 |
| (E1-E4) | SD | .3 | .3 | .3 | .3 | .3 | .3 | .3 |
| | n | 164 | 686 | 776 | 682 | 686 | 476 | 3,364 |
| Training | $\bar{x}$ | 68 | 84 | 91 | 89 | 95 | 96 | 89 |
| Scores | SD | 21 | 22 | 20 | 22 | 15 | 21 | 23 |
| (0-100) | n | 91 | 386 | 387 | 316 | 384 | 212 | 1,882 |
| SQT | $\bar{x}$ | 68 | 85 | 86 | 85 | 86 | 88 | 85 |
| (0-100) | SD | 16 | 13 | 12 | 12 | 12 | 12 | 13 |
| | n | 88 | 348 | 444 | 386 | 476 | 387 | 2,821 |

Note: These n's do not include cases that were high school graduates, females, non-veterans, or early separations for good causes. For this table only, MAP scores were divided into one interval of all sub-passing scores plus five intervals of roughly equal n.

Figure 1

Mean Validities by MAP Score




(46% vs. 54%) in service beyond six months. As for AWOLs (n=649), their mean scores on MAP were close to those for the remaining cases (72.03 vs. 72.19), but their rate of success over one to three years was much lower (.35 vs. .59). The correlation of MAP with tenure for the 2,564 adverse discharges alone was .03.

In the large-fill MOS (combined n=3,270), the rate of six-month success ranged from 69% to 94% (chi square [13 df]=97; p<.001), the median being 87.5%. Mean scores on MAP in these MOS ranged from 71.53 to 73.08 (SD from 5.79 to 7.07). Six-month validities for the large MOS ranged from -.17 to +.31, the difference between the extreme r's being highly significant (z= 3.75; n of 111 and 127; p<.001). The next two most extreme r's (.13 and -.05; n of 297 and 204) were almost significantly different (z=1.93; p<.06).

## DISCUSSION

Taken either datum-by-datum or as a whole, the evidence does not show a strong relationship between scores on operational MAP in FY 81/82 and criteria that are indicators of adaptation. Even though many of the validities are significant at p<01, the proportions of variance accounted for are never as high as .1. Several sources may have contributed to these findings, including problems with the criteria, motivation/faking, and changes in cohorts over time.

First, for over half of the adverse discharges, the cause is not recorded in behavioral terms. Thus adverse discharge occurs for a variety of vaguely identified causes. Also, rates of adverse discharge are subject to a host of influences besides suitability screening (e.g., changes in retention policy at several levels of command; changes in the supply of adaptable youth). These facts work against the reliability of not only the criterion of success vs. adverse discharge, but also the criterion of tenure, for tenure is terminated almost exclusively here by adverse discharge. The only variable which affected validities and rates of success was MOS, which implies that success may be due to an interaction of personal attributes and those of work environment. Recently, research on these latter variables (Olson, Borman, Roberson, & Rose, 1984) has been started at The US Army Research Institute (ARI). Such work may make it possible to improve measures of successful adaptation. Whether adverse discharges during training occur for the same reasons as those after training is also open to question. If not, then those differences may be a further source of unreliability in criteria of success vs. failure.

Secondly, the present cases, being applicants, may have been motivated to answer MAP so as to maximize their chances of enlistment. On the other hand, the development sample in 1977, being in the Army already, may have been willing to answer more candidly. Whether such motivational factors lead to operational faking is now being examined at ARI (Walker, 1984).

Finally, the likelihood of a drift between the 1977 sample and the FY 81/82 sample is confirmed by two sources. First, enlistment standards were much lower in 1977, when ASVAB was misnormed and NHSG were close to 50% of accessions (Grafton, Mitchell, & Wing, 1983). While standards were rising, the recruiting climate was improving, too. Second, the six-month success rates in the development sample (80.2% in 2,280 NHSG of all ages) and the present one (85.5% in 5,941 17-year olds) were significantly different (phi=.067; chi square [1 df]=38.86; $p<.001$). The high success rate in the present cases approaches the 86.4% rate of 8,312 graduates in the 1982 developmental work (Erwin, 1984). These data, along with the above average (for accessions) AFQTs of the present cases suggest that these 17-year olds had been heavily screened before taking MAP.

Previous research on MAP at ARI (Eaton, et al., 1982) gives evidence that is consistent with the hypotheses of faking, cohort drift, and preselection of operational examinees. In that earlier work, 31% of the key development sample failed MAP, compared with 8% of the cases here (phi=.28; chi square [1 df]=813; $p<.001$). The 4.4% of present accessions who had failed MAP may or may not be the same sorts of cases as the 31% failing MAP in the 1977 data. The possibility that the 4.4% may have received particularly heavy screening is suggested by their performance on the criteria, which was higher than expected from the earlier research as well as from its deviation from the trends in outcomes of those who passed MAP (Fig. 1).

Further analyses of the MAP datasets will test these explanations of the present findings. In the meantime, three conclusions are warranted. First, the fact that two-thirds of adverse discharges occur after six months shows that success beyond six months needs to be tracked during key

development to maximize the utility of scoring keys. The present datasets could well be the basis for such work. Second, the variation in success rates across MOS shows that diverse samples are needed for effective key development. Diversity does seem to have been achieved in the latest development work (Erwin, 1984), which gathered data at seven reception stations. Finally, any instrument like MAP needs to be monitored from the time of its implementation. Such monitoring would have the advantage of using data from cases who have the appropriate motivation (i.e., applicants) and would enable periodic updating of keys without special data collection.

## REFERENCES


Atwater, D.C. & Abrahams, N.M. (1983). Adaptability screening: Development and initial validation of the Recruiting Background Questionnaire (RBQ) (NPRDC TR 84-11). San Diego: Navy Personnel Research & Development Center.

Eaton, N.K., Weltin, M., & Wing, H. (1982). Validity of the Military Applicant Profile (MAP) for predicting early attrition in different educational, age, and racial groups (Technical Report 567). Alexandria, VA: US Army Research Institute.

Erwin, F. (1984). Development of the new Military Applicant Profile (MAP) autobiographical questionnaires for use in predicting early Army attrition Alexandria, VA: US Army Research Institute.

Frank, B.A. & Erwin, F.A. (1978). The prediction of early Army attrition through the use of autobiographical information questionnaires (TR-78-A11). Alexandria, VA: US Army Research Institute.

Grafton, F.C., Mitchell, K.J., & Wing, H. (1983). Final status report on the comparability of ASVAB 6/7 and 8/9/10 Aptitude Area Score Scales (Selection and Classification Technical Area Working Paper WP 82-7). Alexandria, VA: US Army Research Institute.

Olson, D.M., Borman, W.C., Roberson,L., & Rose, S. (1984, August). Relationships between scales on an Army Work Environment Questionnaire and measures of performance. Paper presented at the American Psychological Association meetings, Toronto, Ontario, Canada.

Walker, C.B. (1984, September). The fakability of the Army's Military Applicant Profile (MAP). Paper accepted for presentation at the meetings of the Human Resources Management and Organizational Behavior Association, Denver, CO.

# INFLUENCE OF SOLDIERS' EXPERIENCES WITH SUPERVISORS ON PERFORMANCE DURING THE FIRST TOUR

Leonard A. White, Ilene F. Gast,
Helen M. Sperling, and Michael G. Rumsey
U.S. Army Research Institute for the
Behavioral and Social Sciences

November 1984

Presented at the Military Testing Association meeting in
Munich, Germany

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

The statements in this paper are those of the author and do not necessarily express the official positions or policies of the U.S. Army Research Institute or the Department of the Army.

# INFLUENCE OF SOLDIERS' EXPERIENCES WITH SUPERVISORS ON PERFORMANCE DURING THE FIRST TOUR

Leonard A. White, Ilene F. Gast, Helen M. Sperling, and Michael G. Rumsey

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

A large Army project is currently underway to validate new and current predictors of first term soldier performance. However, job performance is not only related to characteristics which are measureable and identifiable prior to enlistment, but is also affected by experiences throughout a soldier's life-cycle in the Army. An understanding of these post-enlistement influences is important to the interpretation of validity coefficients linking pre-enlistment predictors to later job performance. Research on leader effectiveness (e.g., Jacobs, 1970; Yukl & Van Fleet, 1982) suggests that one potent determinant of job performance is a soldier's experiences with his/her superiors. Thus, a research plan was developed to examine leader influences on soldier effectiveness in the context of this larger Army project. As a starting point in this effort, the present investigation was conducted to identify dimensions of leader behavior relevant to the performance of first term Army enlisted. Based on these dimensions, measures will be developed and used in subsequent research to examine leader influences on soldier performance.

During the past 35 years, many different taxonomies of leader behavior have been proposed (see Yukl, 1981, for a review). Each of these research efforts has led to a somewhat different set of categories and there has been little attempt to integrate the findings of various investigations (Yukl & Nemeroff, 1979). Of the categories emerging from this work, the two most widely investigated have been Consideration and Initiating Structure (Stogdill & Coons, 1957). Conceptualizing leader behavior in terms of two dimensions provided a useful starting point for research. However, it is now apparent that this approach fails to isolate important leader behaviors that occur in the work environment (Yukl, 1982). For example, measures of Initiating Structure confound behaviors pertaining to clarification of subordinate roles, use of discipline, monitoring operations, and work planning. Similarly, measures of leader consideration fail to differentiate between performance recognition, providing support, and allowing subordinate participation. In addition, most taxonomies fail to include a number of leader behaviors known to influence work performance such as goal setting and performance modeling.

Recent research by Yukl and his colleagues (e.g., Yukl & Nemeroff, 1979) has attempted to provide a more comprehensive taxonomy of leader behavior which is general enough to apply to different kinds of leaders, yet specific enough to provide relatively homogeneous categories of leader behavior. Based on extensive research (e.g., Yukl & Van Fleet, 1982) with military officers and civilian supervisors, these investigators identified 23 dimensions of leader behavior, which were later collapsed to 12 dimensions. The 12 categories are planning and organizing, monitoring, problem solving, clarifying roles and objectives, motivating task commitment, recognizing and rewarding, developing, informing, consulting and delegating, supporting, harmonizing, and representing. The present research extends the work of Yukl

and his colleagues by using the 12 dimension taxonomy as a basis for classification of Army leader behaviors relevant to the performance of first term enlisted.

## Method

### Collection of critical incidents

A critical incident methodology was used to collect examples of leader behaviors that resulted in effective (or ineffective) soldier performance. Two critical incident workshops were conducted with 80 NCO in five MOS: 05C, 11B, 19E, 63B, and 91B. Participants were trained to write behavioral examples. In each example, NCO were asked to specify the circumstances leading up to the incident, the leader's behavior, and the effect of the leader's actions on soldier performance. When identifying examples, NCO were encouraged to view performance in a broad context to include performance on job-related tasks, motivation, morale, and reenlistment. In the workshops, these NCO generated a total of 474 examples of leader influences on soldier performance. The leaders described in the incidents were primarily NCO.

### Evaluation of Yukl taxonomy

Yukl (personal communication, 1984) indicated that some modification of the 12 dimensions probably would be required to capture the specific influences of Army leaders on subordinates. Thus, the 12 dimensions were evaluated by the authors to ascertain if the Yukl taxonomy omitted any important leader behaviors likely to influence soldier performance. This evaluation was based on a review of the critical incidents collected in this investigation, research on managerial effectiveness, and Army leadership manuals. As a result of this analysis, one category termed "Discipline/ Punishment and the use of Constructive Criticism" was added to yield a total of 13 dimensions. In addition, Yukl's category "Motivating Task Commitment" was expanded to include the motivational functions of positive and negative role models and termed "Leading by Example". Other categories were embellished to provide more specific coverage of the behavioral requirements of Army leaders, particularly NCO.

To determine if the incidents fit into the revised categories, two of the authors classified a sample of 307 incidents using the 13 dimensions. The categories selected by the authors matched for 90% of the incidents. Most incidents were classified into a single category. A secondary category was identified for 47 incidents containing multiple behaviors that involved more than one dimension. All disagreements between the author-raters were resolved through open discussion.

As a check on our classification schema, the incidents were categorized by 31 NCO in MOS 19E who were familiar with Army leadership requirements. The 13 dimension taxonomy was then re-evaluated based on a cross-referencing of results obtained by the authors and NCO raters. Prior to obtaining NCO ratings, each of the 307 incidents was randomly assigned to one of six "Leadership Example Rating Booklets". The examples contained in each booklet were classified independently by groups of 5 or 6 NCO. Raters were asked to place each incident into a single category, but were allowed to use two categories if the incident clearly involved more than one type of

leadership behavior. A miscellaneous category was to be used if the incident did not appear to fit one of the 13 dimensions. After selecting a category, judges indicated the level of effectiveness displayed by the leader in the example. Ratings were made on a nine-point scale ranging from 1 (extremely ineffective) to 9 (extremely effective). Those examples that show good agreement regarding effectiveness ratings and categories may be used as benchmarks defining different effectiveness levels in each category.

## Results

A total of 307 incidents were rated by two of the authors and NCO. Percentage agreement among NCO raters was 60% or better for 225 (73%) of the incidents (Mean = 64% for all incidents). At least 80% agreement was obtained on 146 (48%) of the incidents. The level of agreement fell below 40% for only 9(3%) of the examples. Multiple categories were used infrequently (2% of NCO judgments). For most incidents, a single dimension was identified and no incident was classified in the "miscellaneous" category.

Table 1 presents the frequency and percentage of incidents in each category for the two groups of raters. A category was assumed to represent an

Table 1
Frequency and Percentage of Incidents in Each Category

| Leader Behavior Category | Rater Group: Authors | Rater Group: NCO | Overlapping Classifications |
|---|---|---|---|
| Planning and Organizing | 36 (12%) | 25 ( 8%) | 19 ( 6%) |
| Monitoring | 8 ( 3%) | 10 ( 3%) | 5 ( 2%) |
| Problem Solving | 7 ( 3%) | 2 ( 1%) | 0 ( 0%) |
| Clarifying Roles | 17 ( 5%) | 8 ( 3%) | 3 ( 1%) |
| Leading by Example | 32 (10%) | 27 ( 8%) | 23 ( 7%) |
| Recognizing and Rewarding | 31 (10%) | 32 (10%) | 29 ( 9%) |
| Training and Developing | 54 (18%) | 26 ( 8%) | 25 ( 8%) |
| Informing | 7 ( 3%) | 13 ( 4%) | 6 ( 2%) |
| Delegating/Participation | 8 ( 3%) | 6 ( 2%) | 4 ( 1%) |
| Supporting | 36 (12%) | 37 (12%) | 34 (11%) |
| Disciplining/Punishing | 69 (23%) | 47 (15%) | 46 (15%) |
| Representing | 1 ( 1%) | 2 ( 1%) | 0 ( 0%) |
| Promoting Teamwork | 1 ( 1%) | 2 ( 1%) | 1 ( 1%) |
| No Consensus | -- | 70 | 112 |

Note. Number of incidents=307.

(a) Percentage agreement less than 50%. (b) Includes 70 incidents on which NCO failed to reach consensus plus 52 that NCO and authors put in different categories.

Table 2
Definition of Leader Behaviors

Planning and Organizing. Planning ahead to accomplish mission, using personnel and resources as efficiently as possible. Organizing and scheduling specific tasks in advance, and securing necessary equipment and materiel. Assigning tasks in a fair and reasonable manner.

Monitoring. Keeping informed about subordinates' progress and level of performance, and events that affect mission accomplishment. Being present as needed at work site to monitor performance and check on work progress. Supervising execution of orders. Permitting subordinates to perform work without excessive interference.

Informing. Specifying goals and performance standards to subordinates. Giving orders and directions that let subordinates know how the mission is to be accomplished and what part they will play. Passing information through the chain of command to subordinates, letting subordinates know about decisions, plans and events that affect their work, and disseminating technical information.

Leading by Example. Setting an example for subordinates to follow through one's own appearance, motivation, personal conduct, military bearing, professional competence, and technical/administrative knowledge. Exhibiting bravery. Being willing to share hardships experienced by subordinates.

Recognizing and Rewarding. Praising effective performance and improvements in performance, showing appreciation for special contributions and achievements, and rewarding performance with tangible benefits. Administering rewards fairly, and rewarding those who are deserving.

Training and Developing. Providing skill training or arranging for it to be provided. Providing assistance in a person's professional growth and career development. Giving individualized instruction on specific tasks. Finding opportunities for SM to practice skills learned in training.

Permitting Participation. Giving subordinates who have relevant information and expertise the authority and responsibility for making task decisions. Letting subordinates have some say in decisions which will affect them. Asking for subordinates' ideas on work problems.

Supporting. Showing consideration for an individual's needs and feelings, being supportive, acting friendly, demonstrating or expressing concern for welfare, safety, health and personal well-being of personnel. Assisting individuals in finding solutions to personal problems (e.g., debts).

Disciplining/Punishing and Use of Constructive Criticism. Taking appropriate corrective action when a subordinate violates a rule, disobeys an order, or has consistently poor performance. Applying sanctions fairly and making sure that the punishment given "fits the crime". Criticizing subordinate mistakes in a constructive, calm, and helpful manner.

incident if at least 50% of the raters used the category to describe the leader's behavior. The rank order correlation coefficient between frequency scores in each category was .87, indicating substantial agreement across rating sources. The same category was used by both groups of raters to classify the incidents in 195 (64%) of the examples. Extent of agreement between authors and NCO was also computed for those incidents on which NCO reached consensus. The percentage of identical classifications by the authors and NCO for these 237 examples was 82%. Taken together, these results indicate reasonable agreement across rating sources, keeping in mind that the expected level of agreement based on random assignment to 13 categories is quite low.

The highest levels of agreement were obtained for categories of leading by example, recognizing and rewarding, and supporting. Low levels of agreement were found for dimensions of problem solving, training and developing, clarifying roles and expectations, and representing. Patterns of disagreement among the categories used by the authors and NCO were examined more closely.

Results of this analysis yielded several findings that were used to modify the 13 dimension taxonomy. First, training and developing was probably conceptualized too broadly. Fifteen examples classified by the authors as representing training were placed in other categories by NCO. Second, clarifying roles and expectations overlapped with behaviors concerned with keeping soldiers informed of decisions, plans, and events that affect their work. Third, a close examination of those examples on which NCO failed to reach consensus revealed that the incidents often involved more than one type of leader behavior and could be classified into several categories. Multiple incidents were most evident in the context of providing feedback or punishment. These behaviors were often associated with leader monitoring, training and/or role clarification. Fourth, the categories of problem solving, representing, and teamwork were used infrequently or inconsistently, with low agreement across rating sources for these dimensions.

The modified version of the 13 dimension taxonomy is presented in Table 2. Based on the results of the present research the taxonomy was reduced to nine dimensions. Problem solving was eliminated as a category and informing was combined with clarifying roles and expectations to form a single dimension. Training and developing was narrowed to differentiate this behavioral category from other dimensions. In addition, categories of representing and teamwork were dropped from the taxonomy.

## Discussion

In this exploratory research, critical incidents described by leaders themselves were used to identify relatively homogeneous categories of leader behavior related to soldier performance. A taxomony of leader influences developed in earlier research (Yukl & Van Fleet, 1982) provided a useful framework for the development of this new taxonomy. Nine of the 12 categories identified by Yukl and his colleagues were represented in the final set of 9 dimensions. As in the Yukl categories, the new taxonomy includes most of the broader categories of behavior found to be important determinants of subordinate performance in past leadership research.

Problem solving appeared in the Yukl taxonomy, but did not emerge as a separate category in the present research. The "problem solving situations" appearing in the incidents described leader responses to sudden, short-term "crises" involving first term soldiers, as opposed to long-term activities of strategic planning or policy formation. In selecting a category for these incidents, raters focused on the specific "problem-solving" actions by the leader (e.g., discipline) which impact directly on soldiers. Thus, within the context of the present research, problem solving behaviors may be viewed as antecedent of more observable performance-relevant leader actions represented in the taxonomy.

Representation was used infrequently as a category, a finding reported by Yukl and Van Fleet (1982) in research with military cadets. This may be viewed as surprising in light of evidence (Yukl, 1981) that a leader's success in obtaining support from other units and superiors is related to unit performance. Perhaps asking NCO to write directly to this category or questioning commissioned officers who function as spokespersons for their units would yield more examples of this aspect of leader behavior.

The development of this new taxonomy is the first step in a series of planned research activities. Future research will quantify how the leader activities represented in this new taxonomy relate to the type of power and influence leaders acquire over their subordinates and multiple indices of soldier effectiveness.

## References


Jacobs, T. O. (1970). Leadership and exchange in formal organizations. Alexandria, VA: Human Resources Research Organization.

Stogdill, R. M., & Coons, A. E. (Eds.) (1957). Leader behavior: Its description and measurement (Research Report No. 88). Bureau of Business Research, Ohio State University.

Yukl, G. A. (1981). Leadership in organizations. Englewood Cliffs, NJ: Prentice-Hall.

Yukl, G. A. (1982). Innovations in research on leadership behavior. Paper presented at the meeting of the Eastern Academy of Management, Baltimore, MD.

Yukl, G. A. & Nemeroff, W. F. (1979). Identification and measurement of specific categories of leader behavior: A progress report. In J. G. Hunt & L. L. Larson (Eds.), Crosscurrents in leadership. Carbondale: Southern Illinois University Press.

Yukl, G. A. & Van Fleet, D. D. (1982). Cross-situational, multimethod research on military leader effectiveness. Organizational Behavior and Human Performance, 30, 87-108.

# THE FAKABILITY OF THE ARMY'S MILITARY APPLICANT PROFILE (MAP)

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

February 1985

Presented at the Combined National and Western Region Meeting
of the Association of Human Resources Management and
Organizational Behavior
Denver, Colorado

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

# THE FAKABILITY OF THE ARMY'S MILITARY APPLICANT PROFILE (MAP)

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

Abstract: The sensitivity to deliberate faking of the Army's biodata screen, the Military Applicant Profile, was tested experimentally. Groups of new recruits responded under instructions to answer accurately or to fake in specific ways. The effect of anonymity (vs. giving names) was tested as well. Implications for theory and practice are drawn.

Employers are perennially concerned that prospective employees may falsify their credentials. Falsification may take the form of either exaggerating or concealing qualifications. During times of a general draft, for example, the services try to counter deliberate failure on enlistment screens. With the labor market as it is now, both private employers (Love, 1984) and the services face the problem of applicants giving information about themselves that is unrealistically positive.

The U. S. Army now uses a background questionnaire, the Military Applicant Profile (MAP), to screen male volunteers who have not finished high school. MAP is a multiple-choice questionnaire that identifies applicants who are likely to adapt successfully to Army life. In content, it covers a mixture of topics, including work history, academics, social activities and habits, athletic activity, and expectations of military life. Response choices are scored in terms of weights that are based on observed success rates of soldiers in the past who picked each choice. As a final step in developing the scoring keys, adjustments were made in the empirically based response weights to reduce the effect on scores of picking the most socially attractive answers. MAP has been validated against a criterion of completion of the first 180 days of service versus involuntary discharge for reasons of conduct (Eaton, Weltin, & Wing, 1982). Recently new forms of the instrument have been developed for use with females and high-school graduates as well as non-graduate males (Erwin, 1984).

The present research addresses two questions: can deliberate faking on MAP affect the way people respond? And can such faking affect scores on the scoring key? These questions are links in a chain of inquiry to determine whether faking is undermining the validity of the instrument in its operational use.

The opinions in this paper are those of the author and do not necessarily reflect the official positions or policies of the U. S. Army Research Institute or the Department of the Army.

[1] 5001 Eisenhower Avenue, Alexandria, VA, 22303. 202-274-8275
Areas of interest: HRD, HRM.

In the present research, naming (vs. answering anonymously) is varied to test the hypothesis that asking for people's names on a self-report instrument is equivalent to asking them to try to make themselves look good. That hypothesis is supported by two experiments. In the first (Haymaker & Erwin, 1980), Army recruits took MAP for attribution under instructions to answer accurately, then retook it under instructions to fake good. Although scores of some items changed significantly, total scores on the instrument did not. Those results are consistent with the notion that faking on the first administration had already raised scores as high as it could. The second experiment (Atwater, 1980) found that Navy recruits who answered the Recruit Background Questionnaire (RBQ) for attribution scored slightly but significantly higher than other recruits who answered anonymously.

The naming hypothesis is important because biodata screens are validated under the condition of responding for attribution. Since instruments like MAP and RBQ have proven usefully predictive, then that predictive power may occur despite name-induced faking. If so, then there may be little additional effect that other sources of faking could have on scores.

Regarding construct validity, the naming hypothesis bears on the question of the meaning of scores on such instruments. Do they indicate candid histories, self-serving performances, or something else?

## Method

### Instruments

A 112-item questionnaire was prepared consisting of the 52 scored items from one form of the operational MAP plus 60 other items from its forthcoming edition. The 60 were chosen for their suspected vulnerability to faking.

### Respondents

Participants were 1836 new recruits at three Army posts during April and May, 1984. These recruits were in their first three days of service at in-processing sites called Reception Stations. In composition the sample was 88% males, 86% high-school graduates or above, and 66% Regular Army (vs. 34% National Guard and Reserves). The mean age was 20, the median being 19. Participants came from at least 147 of the entry-level military occupational specialties, 101 persons being by far the most from any one specialty.

## Procedure

The author administered the instrument in large testing rooms to 21 groups which had a median N of 65 (range: 8 to 226). In total, the time for instructions and responding was about 50 minutes, with about 90% of the participants finishing. Respondents answered on a machine scorable answer sheet.

All groups were told that this data collection was for research purposes only and that they were to play a role while answering the questionnaire: the role of civilian applying for military service. The instructions diverged into four treatments at this point. As a baseline, one group was told to answer as accurately and honestly as possible. They were told that this research is for developing better ways of scoring the new MAP.

In contrast, the other three groups were told to try to make their pretended applicant look good or bad. In two "fake good" conditions, one group was to try to answer so as to look as attractive as possible to the Army (fake unrestrainedly good), the other to look good, but not so much so that the pretense would be obvious (fake discreetly good). The final group was to answer so as to make themselves look unattractive to the Army, but believably so (fake discreetly bad). To justify the faking instructions, these groups were told that this experiment was for developing ways to identify and counter faking on the new MAP. The researcher answered questions about the tactics or content of faking by saying to pick the answer that the recruit thought the Army *would* or *would not* want to hear.

Anonymity (vs. responding for attribution) was crossed with the four treatments by asking some groups for names and social security numbers on the answer sheets.

## Analyses

The questionnaire items were analyzed in terms of two metrics, the raw item scales (response A=1, B=2, etc.) and the empirically keyed response weights. The latter were available for the 52 scored items from the operational form of MAP. At the level of the individual item, the effect of instructions on raw scales was tested with chi square, while the effect on the scoring weights was tested with t-tests. The two scales serve to answer different questions: the raw scales show whether the treatments are associated with different patterns of responses, while the scoring weights show whether instructions affect the output of the scoring key.

Statistical significance being sensitive to sample size, Cramer's V was also used on the raw response scales. Values of V are affected by the degrees of freedom, so it gives lower bound estimates of strength of effect, on a scale of 0 to 1.

To produce a total score for any set of MAP items, it is not meaningful to deal with the raw item scales. The items differ in numbers of choices; some are entirely categorical; and many have only partly ordered response choices. To get a total score on MAP for a respondent, the keyed item scores, all of which have a range of 0 to 2, are simply summed. Analyses of variance were run to test the effects of the manipulations on total MAP scores.

## Results

The sample for analysis included 1788 recruits, after 48 had been excluded for answering less than 60 ofthe questions or defacing the answer sheets. Before further selection of cases, the Ns in the four treatments were 631, 481, 525, and 151 for fake unrestrainedly and discreetly good, controls, and fake bad, respectively.

For the first outcome measure which was analyzed, the raw item scales, Table 1 shows the numbers of items out of 112 which had significantly different distributions ($p<.05$) in selected pairs of treatments. It is apparent that the various instructions changed these distributions for most of the individual items. Table 1 also gives the median Cramer's V for the items in each condition which had significant chi squares. The strongest departures from the controls were in the fake unrestrainedly good and fake discreetly bad conditions.

For the second item metric, the empirically weighted responses, t-tests were run on each of the 52 keyed items. The predicted magnitudes of scores were as follows: fake unrestrainedly good>fake discreetly good>control>fake discreetly bad. However, the results did not consistently support these expectations in the first three treatments. In comparisons of the first two conditions, 37 of the 52 items had means that were significantly ($p<.05$) different, but for 26 of those, the discreetly good group scored higher. The maximum strength of effect (omega square) for that set of tests was .0695. Comparing unrestrained faking good with the controls, 34 items differed significantly, but 23 of those favored the controls.

On the other hand, while only 23 items produced significant differences between the controls and the discreetly good group, 16 of those means were higher for the fakers. The greatest strength of effect here (omega square) was .0488. Finally, the fake bad condition gave consistent results: 44 items differed significantly, 38 of those favoring the controls, and strength of effect running as high as .21.

Total scores on the operational MAP were analyzed in a Treatments by Anonymity analysis of variance. Only the 1768 recruits who answered 47 or more of the 52 keyed items (i.e., 90% or more of the items) were included. Scores of those who answered 47 to 51 of these were proportionally adjusted to a

base of 52. In Table 2, cell means from this analysis are shown. The different instructions had a strong effect ($F(3,1760)=408.6$; $p<.0001$), even when the faking bad condition was not included ($F(2,1620)=44.5$; $p<.001$). Naming and its interaction with treatments had no discernible effect. Each faking condition was significantly different from the control ($p<.001$), but the two fake good conditions did not differ significantly from each other.

The effects of giving or not giving names on distributions of raw item responses were tested only on the responses of the controls because that condition resembles preinduction testing most closely. For this analysis there were 380 anonymous and 143 named respondents. On only six of the 112 items did the naming manipulation have a significant effect.

## Discussion

Instructions to fake deliberately in answering the Army's MAP do influence the distributions of response choices. The effects are frequent and strong, judging from the number of items showing significant chi-squares and from the values of the Cramer's statistic. The strongest departures from the control condition are in the two groups which are conceptually most different from it: faking to look extremely good and faking in the direction opposite that of everyday impression management. Table 1 shows that the nominal difference between faking good unrestrainedly and discreetly is psychologically real to service-age youth. How often they can and will make that distinction in the operational setting is open to question.

From the absence of effects due to giving names, we draw two conclusions: asking for names did not have the same effect as instructions to fake good; and, the controls were probably answering validly even when not protected by the safety of anonymity. Although our hypothesis on the effects of giving names is not supported, the results confirm the harmlessness of asking for names in collecting data for the announced purpose of research, at least with an item pool like MAP's, that excludes intrusive or sensitive questions.

The influence of experimental faking on the operational scoring system is significant at the item and total test level, but the effects are large and consistent only in the faking bad condition. As noted above, the empirically derived scoring keys were originally designed to counter faking good. The data here imply that the adjustment keeps the effects of that tactic small. But MAP is very vulnerable to faking in the direction it was not keyed to resist. Research under different conditions of the labor market would be needed to produce a biodata instrument which could resist faking bad by draftees.

One difference in the information in the raw item scales and in the scoring weights is that the fake discreetly good condition was most similar to the controls in the former but most similar to the extreme fake good condition in the latter. That is, discreet faking good is not as obvious (in terms of raw response distributions) as extreme faking, but it affects the keyed scores just as much. Both of the fake good treatments produced gains of about four points out of a possible 104 on the total test score. That gain replicates the size of effect of fake good instructions in a pilot of this experiment on 500 recruits this past spring.

What is the import of this four-point mean difference for the operational validity of MAP? That is not clear from these data because we do not yet know the frequency of faking in operational testing. The means of the control and fake good groups are less than one standard deviation (of the scores of individuals) apart. From archival data we know that only 14% of applicants who took the MAP in fiscal years 1982 and 83 scored in the four-point interval above the cut score. Nevertheless, such large numbers of applicants will take MAP when its use is extended to all applicants that it may prove appropriate to try to counter faking.

The research literature on faking finds that practically any measure can be deliberately faked, all the way from involuntary bodily reactions to lie-detection scales. A finding that a measure <u>can</u> be faked does not automatically mean that it <u>is</u> faked in an applied setting. Thus the finding that MAP is somewhat fakable is necessary to show the potential for operational faking, but not sufficient to show the reality. Further work in this research program will develop fake-detection keys and then apply them to archival data on applicants in 1981/82. Only if the applicants who showed a faking pattern of responses had sub-par records of later performance in the Army, and only if these cases occurred in any sizable numbers, will faking to look good need remedying.

Further data collection in the present experiment is under way to gather enough numbers for cross-validating fake detection keys and to test the resistance to faking of new scoring keys.

Table 1
Numbers of items out of 112 having significantly different distributions of responses in selected pairs of treatments (Chi square tests; alpha=.05)

| Comparison | Fake Very vs Discreetly Good | Fake Very Good vs Control | Control vs Fake Discreetly Good | Fake Discreetly Bad vs Control |
|---|---|---|---|---|
| No. of sig. differences | 103 | 110 | 63 | 108 |
| Median Cramer's V of the sig. items | .20-.24 | .30-.34 | .10-.14 | .40-.44 |
| Ns in each Treatment | 628/479 | 628/518 | 518/479 | 151/518 |

Table 2
Mean total scores, standard deviations, and cell Ns by treatments and naming conditions

| Treatment | Mean | Standard Deviation | N |
|---|---|---|---|
| Fake unrestrainedly good | | | |
| Anonymous | 68.5 | 7.7 | 285 |
| Named | 68.9 | 6.3 | 343 |
| Fake discreetly good | | | |
| Anonymous | 68.6 | 6.9 | 264 |
| Named | 67.9 | 6.7 | 215 |
| Control | | | |
| Anonymous | 64.3 | 8.5 | 377 |
| Named | 65.3 | 7.1 | 142 |
| Fake discreetly bad | | | |
| Anonymous | 41.0 | 13.3 | 103 |
| Named | 44.4 | 13.6 | 39 |

## References


Atwater, D. C. (1980). Faking of an empirically keyed biodata questionnaire. Paper presented at the meeting of the Western Psychological Association.

Eaton, N. K, Weltin, M., & Wing, H. (1982). Validity of the Military Applicant Profile (MAP) for predicting early attrition in different educational, age, and racial groups (Technical Report 567). Alexandria, VA: U. S. Army Research Institute.

Erwin, F. W. (1984). Development of new Military Applicant Profile biographical questionnaires for use in predicting early Army attrition. Unpublished manuscript.

Frank, B. A. & Erwin, F. W. (1978). The prediction of early Army attrition through the use of autobiographical information questionnaires (Technical Report No. TR-78-A11). Alexandria, VA: Army Research Institute.

Haymaker, J. C. & Erwin, F. W. (1980). Investigation of applicant responses and falsification detection procedures for the Military Applicant Profile (Final Project Report, Work Unit No. DA644520). Alexandria, VA: Army Research Institute.

Love, A. M. (1984). Specializing in untangling the webs of resume deceits. Philadelphia Inquirer, 21 May 1984, 4C.

Walker, C. B. (1984). Validation of the Army's Military Applicant Profile (MAP) against an expanded criterion space. In B. Means (Chair), Recent developments in military suitability research. Symposium held at the meeting of the Military Testing Association, Munich FRG.

# ASSESSING THE UTILITY OF A PERSONNEL/CLASSIFICATION SYSTEM

Robert Sadacca John P. Campbell
Human Resources Research Organization

March 1985

Presented at the meeting of the Southeastern Psychological Association at Atlanta, Georgia

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

# ASSESSING THE UTILITY OF A PERSONNEL/CLASSIFICATION SYSTEM[1]

## Introduction

A major issue in the development and evaluation of a personnel selection or classification system concerns how to assess the net gain to the organization from using the new system vs. not using it. It is a natural question for management to ask and it has been a major issue in the professional personnel research literature for quite some time.

Historically, the issue has been addressed by casting it into a decision-making framework and treating it as a decision-making problem. As a result, previous research and theory on decision making from a variety of disciplines become relevant. Once these steps are taken, the two principal questions then become: (1) how can the payoff from a particular course of action be evaluated, and/or (2) how can the relative payoff from different courses of action be compared?

To answer such questions at least three major things are needed: (1) a model that portrays the relevant parameters in the decision-making process and specifies how they are interrelated, (2) a metric that can be used to represent the value of the outcomes that result from a particular course of action, and (3) a method for estimating the parameters of the model in the appropriate metric.

We know a fair amount about modeling personnel selection decisions (e.g., Cronbach & Gleser, 1965) and somewhat less, but still quite a bit, about modeling personnel classification decisions (e.g., Rulon, Tiedeman, Tatsuoka, & Langmuir, 1967). A great deal of effort by psychometricians and industrial psychologists has been put into the development and refinement of such models (cf. Cascio, 1982a). We are much less clear as to the metric in which the outcomes of a personnel selection or classification decision should be expressed.

## The Utility Issue in Industrial Psychology

Although the following steps have not occurred in a perfect chronological order, the progression of attempts by psychometricians and personnel researchers to portray the benefits of selection and classification has been something like the following.

The validity coefficient, in the form of the product moment correlation between a predictor composite and a criterion composite, is the classic method by which the value of a selection program is represented. However, as is widely acknowledged, the correlation coefficient is a difficult metric to

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

interpret. Early on, a number of transformations such as the coefficient of determination ($r^2_{xy}$), the index of forecasting efficiency ($1-\sqrt{1-r^2_{xy}}$), and the standard error of prediction ($S_y \sqrt{1-r^2_{xy}}$) were suggested and found wanting. They still depended very heavily on the correlation coefficient, itself, and cannot be interpreted directly in terms of benefits from decision making.

A more useful kind of transformation is represented by the various ways of using the bivariate distribution to construct decision tables. The Taylor-Russell tables (Taylor & Russell, 1958) are examples. With these transformations, the metric becomes the proportion of correct predictions that are made by one selection method vs. another. One benefit of looking at selection payoff in terms of decision accuracy is that it illustrates quite clearly how even a small relationship between predictor and criterion can produce significant gains in the number of successful people selected if the selection ratio is very low and/or the variability in performance is high (e.g., base success/failure rate = .50). However, to express the value of selection in these terms the organization is required to define specific criterion categories (e.g., successful vs. unsuccessful performance) and to view all the outcomes in a particular category as being equally valuable.

A new dimension was added by the classic work of Brogden (1946) who showed that if both the predictor and criterion measures had interval properties and if the relationship between them was linear, then the correlation coefficient is linearly related to the gain in performance in the selected group. Further, the gain, in standard criterion units, that will result from selection can be estimated using existing prediction (i.e., decision) models if a cutting score is set on the predictor. Brogden also argued that a desirable metric for performance and performance gain would be to determine the dollar value of variability in performance. It remained for Cronbach and Gleser (1965) to add the consideration of selection costs and to portray the utility of selection benefits in terms of the dollar value of performance increases minus the costs of selection. Cronbach and Gleser also elaborated the utility formulation to include more complex selection modes (e.g., multiple hurdles) and made an attempt to formulate classification decisions in utility theory terms.

The application of this kind of utility/decision theory to selection and classification problems was hampered by the difficulty of estimating the variability of performance in dollars, which is a major parameter in the model. Recently, Schmidt and Hunter (1979) proposed a rather simple solution in which supervisors are used as judges to scale individual performance in dollar terms via a magnitude estimation technique. Judges are asked to estimate the dollar payoff of performance at the 50th percentile and the 85th percentile for people in the job in question. The difference between the two estimates is taken as the standard deviation of individual performance in dollar terms ($SD_y$). So far, not much attention has been paid to the basis on which supervisors make such judgments although the value for $SD_y$ is frequently between 40 and 60 percent of the annual salary for the position.

Cascio (1982b) has proposed another technique for estimating $SD_y$ in dollars that also uses expert judgment and is tied explicitly to salary. Job analysis is used to determine the major task components of a job, their relative importance is determined by expert judgment, and a magnitude estimation

technique is used to rate every person's performance on each task factor. Average total salary is apportioned to each factor in accordance with its importance weight. Average performance is set equal to 1.0 and the resulting scale is multiplied by the proportion of salary designated for that factor. Performance differences have thus been converted to a dollar metric and the standard deviation of the aggregate differences are put into the Cronbach and Gleser equation.

## Utility Judgments in the Military Context

Two principal factors make it difficult to apply the previous work on utility metrics and utility estimation to the Army context. First, compensation practices in the Army vs. the civilian sector are quite different. Salaries to not differ by MOS and thus cannot be used as an index of the job's relative worth to the organization. Second, the Army is not in business to provide products or services so as to maximize profit. Its overall mission is to be prepared to defend the U.S. against military threats that everyone hopes will never come. It is difficult to put a monetary value on success or failure or to even think of the utility of jobs in terms of their monetary benefit. Dollars may not be an appropriate metric with which to evaluate a new classification system aimed at maximizing preparedness for catastrophic events. However, resources are not unlimited and choices among alternative personnel practices will be made whether or not there is an explicit utility metric on which to make comparisons. One operational answer to the problem is the system currently in use in the U.S. Air Force.

### The Air Force Procedure

Entry level assignments in the Air Force are made by the PROMIS selection and classification system (Ward, Haney, Hendrix, & Pina, 1978). In very brief terms, the individual assignment is a function of the following five parameters:

1) The level of predicted training success using the Armed Services Vocational Aptitude Battery (ASVAB) and other applicant information as predictors.

2) The individual's job preferences.

3) The rate at which the targeted quota for a job is currently being filled.

4) The rate at which the minority group targets for each job are being filled.

5) The scaled importance value of each job holder aptitude level x job difficulty combination.

It is this last parameter that serves as the analog for a utility metric in the Air Force system. Previous scaling research using expert judges has produced an overall scale value for the relative importance of each combination of job difficulty (as determined by expert judgment) and the aptitude

level of a job holder as determined by ASVAB scores. In general, the greater the job difficulty or the higher the aptitude level of the individual, the higher the value of that personnel assignment. However, the prediction surface that relates the aptitude level/difficulty level combination to assignment value is not a linear plane.

The approach of Project A[2] to the problem is similar but different. Instead of scaling the relative importance of job difficulty x aptitude level combinations, the focus of Project A is on assessing the differential value, or payoff, from MOS x predicted performance level combinations.

## Specific Utility Issues for Project A

The overall objective of Project A is to produce the information necessary to develop a functional personnel classification system for all enlisted personnel. The objectives of Project B, a concurrent, related effort also under the direction of the U.S. Army Research Institute, are to develop the necessary algorithms for relating labor supply forecasts, applicant information, and forecasts of system needs in an assignment system that uses Project A data in an optimal fashion.

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

For Project A the problem of estimating such utility values breaks down into a number of specific questions.

First, there is the matter of how performance levels should be defined. Should it be in terms of some general performance dimension that is left unspecified and defined only in terms of relative rank (e.g., percentiles)? Should a general performance dimension be explicitly defined perhaps with behavioral anchors developed via critical incident methodology? Should individual performance components be defined and then explicitly weighted for combination into a total score? All of these are possibilities and a specific research question concerns how performance levels should be defined and described in the MOS x performance level combinations.

A second specific question concerns what is the most appropriate metric for describing the relative value, or utility, of differential assignments

---

[2] Project A, Improving the Selection, Classification, and Utilization of Army Enlisted Personnel, is a nine-year, large scale program designed to provide the information and procedures required to enlist, allocate, and retain the most qualified soldiers. It is being conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences.

across MOS/performance level combinations. Previous work in a selection context in personnel psychology has appealed almost exclusively to a dollar metric and has tried to estimate the variability in payoff from people at different performance levels in dollar terms. However, estimating differential payoff from a system wide classification system remains unexplored. Since the dollar metric may not be appropriate for the Army context and because there is little previous work on applying utility theory to personnel classification, the metric question for Project A is a very difficult one. It suggests an exploratory approach.

Assuming the question of the metric is resolved, the specific method(s) to be used for estimating differential assignment utility in the appropriate metric must then be considered. Only two options seem even possible. First, it might be possible to relate the performance of individuals or units to some kind of "bottom line" measure that Army management would consider an appropriate metric. For example, realistic field exercises could be used to determine the relationships of individual performance measures to the performance of a unit in a simulated engagement. The difficulties with this approach revolve around the expense of collecting such data, the necessity of having such exercises for each MOS, and the necessity for equating scores in some way across MOS. A second alternative, which could be combined with the first, is to appeal to scaling technology and to use expert judges to estimate the relative value of differential personnel assignments. There are a variety of scaling models and scaling techniques from which to choose and a major difficulty would be in choosing the procedure which is feasible, makes the best use of the information held by the judges, and provides sufficient internal validity information to generate confidence and acceptability for the scale values.

Since the above questions are difficult ones and have been largely unresearched in the past, the plan that was developed for addressing them is exploratory in nature. It tries to proceed from a very broad consideration of all possible issues to a focus on a procedure that is valid, feasible, and acceptable to the Army.

## General Procedure

The general procedure for arriving at estimates of assignment utilities for MOS x performance level combinations will involve three phases, the first of which is complete.

Phase one consisted of a series of seven small group workshops with Army officers. The workshops were designed to explore a number of issues pertaining to utility, utility metrics, utility estimation, and the definition of performance levels. Each workshop was divided into a period for trying out prototypic judgment tasks and a period for open-ended discussion of issues.

Although the atmosphere was informal and the participants were free to bring up any questions or issues they wished, the questions that were used to guide the discussions were the following.

1. How shall measures of performance be weighted and overall performance defined?

2. What kinds of scaling judgments can officers reasonably be asked to make?

3. Are there major scenario effects on performance factor weights and utility judgments?

4. In what metric should the utility of enlisted personnel assignments be expressed?

5. What is the form of the relationship between performance and utility within MOS?
6. Who will make the best judges for the final scaling?

The specific judgment tasks that were tried out in phase one will be discussed in more detail in the next section of this paper; however, their general nature was as follows:

1. Assignment of importance weights to performance factors.

2. Rank ordering overall utility of MOS x performance level combinations when performance was defined in percentile terms.

3. Ratio judgments of comparative utility for different MOS x performance level combinations.

The specific reactions of each participant to the sample scaling tasks were also used as items for general discussion.

Phase two will consist of another series of workshops devoted to a more focused and in-depth exploration of the utility issues and estimation problems identified in phase one. They will include obtaining reactions from the participants to descriptions of hypothetical results obtained with a new classification system when the payoff from using the system is expressed in a variety of ways. The final three workshops in the series will be devoted to a systematic tryout of what seems then to be the two or three best approaches to utility scaling.

Phase three entails the collection of the utility data that will actually be used with the 83/84 cohort validation data to develop the techniques for estimating overall classification validity.

## The Exploratory Utility Workshops

The preparation of the utility judgment procedures, the conduct of the exploratory workshops, the analyses of the judgment data, and the subsequent preparations for the next workshop can perhaps best be viewed as a research process. There was no vigorous testing of hypotheses, no experimental design or testing for statistical significance. If something didn't seem to work it was dropped or modified; if something else occurred to one of the present writers or was suggested by one of the workshop participants it was tried out. We were essentially trying to figure out what could possibly be done, before worrying about how to most effectively do it.

**Workshop 1**

A critical initial concern was whether Army officers would be willing to make evaluative judgments comparing the utility of enlisted soldiers in different military occupational specialities (MOS). Officers might, for example, argue that all military jobs are essential, and that it does not make sense to say that the soldier who transports the ammunition has any less utility than the soldier who fires the weapon, or the soldier who treats the wounded, or the soldier who feeds the troops.

Assuming that officers would be willing to assign different utility values to different enlisted MOS/performance level combinations, another critical concern was what military situation or scenario should be used as the context in which the judgments of utility are made. It seemed very reasonable to believe that the utility to the Army of different military jobs and performance levels within those jobs, would vary as a function of the stipulated military situation. Infantrymen or armor crewmen at most if not all levels of performance, for example, would most likely be assigned higher utility values under a wartime scenario than under a peacetime one. Likewise, differences in utility values of different MOS/performance level combinations could well exist from one wartime scenario to another, e.g., armor crewmen might be judged to have relatively less utility in a jungle war than in a European conflict.

A third concern centered on what considerations enter into judgments of utility made by Army officers. When evaluating a soldier's utility, what contributions to mission accomplishment are the officers emphasizing? For example, are they thinking more in terms of inflicting damage on the enemy or survival of their units?

To get an initial understanding of the first two issues, it was decided not to provide any military context for making the utility judgments to officers attending the first workshop (six field grade officers from the Army Research Institute, the organization sponsoring Project A research). We were interested in finding out whether they would evoke their own military context for the judgments, and if so, what context would they choose. As we were also concerned with the reasonableness of the utility assignment task from the officers' point of view, we also decided to keep the scaling task at the ordinal level, that is, to only ask for a rank ordering of MOS performance/level combinations rather than for more sophisticated judgments that could yield an interval or ratio utility scale.

Subsequently, after a brief introduction to Project A and a discussion of the concept of job performance utility, the six officers were given the task of rank ordering a set of 57 enlisted MOS/performance level combinations. Exhibit 1 gives the directions used for the utility rank ordering and an example of the stimulus cards that contained the MOS/performance level combinations. The officers were also given a separate

listing of the 19 Army Military Occupational Specialty (MOS)[3] summary job descriptions to facilitate their judgments (Exhibit 2). Although administered in a group setting, the officers independently performed the rank ordering.

Perhaps the most important result of this first attempt at obtaining utility values of Army MOS was the one that was immediately apparent -- the officers were willing to do the task. They did not argue that it was an unreasonable one as we had feared they might. They seemed to undertake the task quite seriously and carefully.

Another significant result emerged in the post task discussion: Each of the six officers had independently chosen the same scenario -- that of a European war -- as the context in which they had rank ordered the utility of the MOS/performance level combinations. In the discussion period, the officers expressed the opinion that the Army's principal current mission is to ready itself for such a possibility. They agreed that had they used a peacetime or a different wartime context, that their utility rankings would most likely have been different. But they felt even if we used a peacetime scenario it should be one that emphasized training and other readiness activities geared toward the outbreak of hostilities in Europe.

The rank order intercorrelations among the officers were computed across the 57 MOS/performance level combinations. These correlations ranged from .29 to .90 with an average of .69. These results were heartening, since they indicated that quite reliable (.95 or above) average utility ranks could be obtained by using 10 or more judges. The results also indicated that there may be a fairly common frame of reference among Army officers in their evaluation of MOS/performance level utilities.

The officers were next asked to evaluate the relative priority of eight[4] outcomes of a military engagement that could result from effective performance of enlisted personnel in that situation. (Table 1 lists the outcomes used.) The military scenario (chosen before the workshop began) was one describing the outbreak of hostilities in Europe that had been used previously in other Project A activities[5]. It is shown in Exhibit 3.

Five officers (one officer had to leave) evaluated the relative priority of the outcomes using two different scaling techniques in counterbalanced order. In one method, they first rank ordered the eight outcomes. Then assigning ten points to the lowest ranked outcome, they assigned points to

---

[3] The 19 MOS constitute the sample MOS selected by Project A to be representative of Army jobs. For a description of how this sample was drawn see Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report (Research Report 1347). Alexandria, VA: HumRRO, AIR, PDRI, and ARI. October 1983.

[4] The eight outcomes were chosen by the writers without regard to any official Army doctrine.

[5] The chosen scenario was used in the evaluation of the importance of specific job-related tasks for various MOS.

Table 1

Outcome Importance for a Wartime Scenario
(n = 5 Field Grade Ofdficers, Workshop 1)

| Outcome | Mean Rank | Mean Scale* |
|---|---|---|
| Increased force survival | 2.2 | 124 |
| Enhanced readiness | 3.0 | 114 |
| Enhanced efficiency or cost-effectiveness | 7.8 | 18 |
| Enhanced mobility and fire power | 2.6 | 108 |
| Enhanced physical and psychological well-being | 5.4 | 61 |
| Increased local civilian cooperation and support | 7.2 | 30 |
| Decreased capability and performance of enemy units | 3.0 | 100 |
| Enhanced performance of supporting Army units | 4.8 | 85 |

*Scale: Assign 10 points to lowest ranked outcome.

the remaining outcomes in accordance with the perceived ratio of their importance to the lowest ranked outcome. In the other scaling method, the officers were presented the eight outcomes in a paired comparison format -- for each possible pair of outcomes their task was to divide 100 points between the outcomes in a manner that reflected the outcomes' relative importance in the given military situation.

Our primary interest in trying out these two scaling techniques was to obtain the officers' reactions to them. They distinctly did not like the paired comparison format, feeling that it was like a test of their consistency in assigning importance points. (Each of the 28 possible pairs was presented on a separate card and the officers had been instructed to rate the pairs one by one without going back over their earlier judgments).

Although the officers did not express any strong negative reactions to the scaling method that involved assigning ten points to the lowest ranked outcome, several of the officers did indicate they would have liked to assign zero points to one or more of the outcomes. Moreover, examination of the distribution of points across the eight outcomes indicated wide between judge variability. The mean number of points assigned the 8 outcomes ranged from 59 for one judge to 119 for another. The standard deviation of points assigned the eight outcomes ranged from 30 to 86. We decided to try out a scaling method that might better control for interjudge differences in assigned points in the next workshops.

The mean rank order assigned the eight outcomes are given in Table 1. These rankings are of interest in two regards. First, they show the relative emphasis put by the five officers on force survival in the fluid battlefield situation described in the wartime scenario used. In the following discussion period, the officers emphasized the importance of keeping units intact and ready to fight in such a war. Reducing enemy capacity, although ranked high, had lower overall priority than unit survival.

Of equal interest was the low rank given to the outcome, enhanced efficiency or cost-effectiveness. This outcome was ranked last by four of the officers, and next-to-last by the fifth officer. In the discussion period, some of the officers indicated that in their opinion dollar cost considerations had no place on a battlefield, that losing or even winning a war could not be evaluated in dollar terms. They further indicated that the costs of training and equipping soldiers did not enter into their MOS/performance level utility rankings.

In response to the question whether judges should evaluate MOS/performance levels against separate utility dimensions, the officers expressed a clear preference for making one overall utility rating. They also felt that the description of the MOS/performance levels should be kept general rather than made more specific.

## Workshops 2 and 3

As the second and third workshops were scheduled back-to-back on successive days, we planned to use the same stimulus materials and judgmental tasks in both workshops. However, discussions with the second workshop officers led to changes in the procedures used the next day.

One such change involved the scenario used to describe the military context for the utility judgments. For the second workshop, we had decided to use the same wartime scenario that was used as the context for the outcome evaluations in the first workshop (see Exhibit 3). Discussions with the six field grade officers in the second workshop indicated, however, that their utility ratings might well have been influenced by the type of unit to which they imagined themselves assigned. Furthermore, they might have been responding differentially to the "rugged, hilly and wooded" terrain description. One officer, for example, reported that he had downgraded the utility of armor crewmen because of the more limited use of tanks in that setting, while other officers reported that they had nevertheless assigned very high utility values to the armor crewman MOS. The officers suggested keeping the scenario(s) free of specific details that would favor one MOS at the expense of another. The references in the scenario to the specific terrain and weather conditions were therefore deleted from the wartime scenario used in the third and subsequent workshops (see Exhibit 3). Moreover, the military unit of concern was made the entire Corps, rather than an unspecified unit within the Corps.

Another change that was made in the evening between the second and third workshops was the directions for scaling the relative importance of different types of Project A criterion measures. The method used in the second workshop entailed dividing 1,000 points among the criterion measures to be weighted to form a performance composite. As the officers in the second workshop were highly critical of this method (they felt too much time was spent trying to make the points add up properly), the method used in the third workshop entailed assigning 100 points to the criterion measure ranked most important and weighting the remaining criterion measures accordingly. In comparison to the method used in the first workshop (assigning 10 to the lowest ranked factor), this method allows the judges to assign zero weights.

In both the 1000-point and 100-point methods, the officers first rank ordered component measures that could be used in deriving Project A overall performance scores. They then applied the assigned scaling technique. Three sets of component measures were scaled (see Exhibit 4). The first set consisted of the principal different types of measures being developed by Project A. These included hands-on performance measures, job knowledge tests, supervisory, peer and self-ratings, and administrative indexes derived from the soldiers' official personnel folder (201 file). The second set of component measures consisted of the 11 behaviorally anchored (BARS) scales being tried out for obtaining ratings of general soldiering performance in Project A. The third set of component measures consisted of 6 administrative measures derived from prior Project A research on first-tour soldier 201 files, e.g., number of letters and certificates of commendation received, whether the soldier was eligible for reenlistment, and number of disciplinary actions taken against the soldier.

Table 2 presents the correlations obtained between the mean rankings and scale values of the sets of component measures obtained from the six officers in Workshop 2 and the seven officers in Workshop 3. The table also presents the average of the interjudge correlations taken across the different types of measures. The average interjudge correlations and correlations among the mean rank and scale values are consistently higher for the third workshop than for the second. This may reflect the general lack of consensus that

Table 2

Correlations among Types of
Performance Factor Weighting Methods
(n = 6, Workship 2; n = 7, Workshop 3)

| | Type of Measure (10) | (1) | (2) | (3) | Average Interjudge Correlation |
|---|---|---|---|---|---|
| (1) | Workshop 2 rankings | -- | | | .25 |
| (2) | Workshop 3 rankings | .79 | -- | | .80 |
| (3) | Workship 2 scale: 1000* | -.87 | -.80 | -- | .22 |
| (4) | Workshop 3 scale: 100** | -.72 | -.97 | .83 | .80 |
| | BARS Scales (11) | | | | |
| (1) | Workshop 2 rankings | -- | | | .31 |
| (2) | Workshop 3 rankings | .80 | -- | | .52 |
| (3) | Workshop 2 scale: 1000* | -.87 | -.68 | -- | .23 |
| (4) | Workshop 3 scale: 100** | -.80 | -.98 | .72 | .43 |
| | 201 File Measures (6) | | | | |
| (1) | Workshop 2 rankings | -- | | | -.14 |
| (2) | Workshop 3 rankings | .52 | -- | | .11 |
| (3) | Workshop 2 scale: 1000* | -.77 | -.68 | -- | -.03 |
| (4) | Workshop 3 scale: 100** | -.40 | -.94 | .53 | .14 |

characterized the second workshop in the discussion periods. The participants disagreed considerably among themselves on a number of the issues that were brought up. The lower interjudge correlations obtained for the 1000-point scaling method vs. the 100-point scaling method may thus be a function of genuine disagreement among Workshop 2 officers, the difficulty of the 1000-point method, or both.

In the discussion of the scaling tasks both groups of officers indicated that it was difficult to assign weights to the 11 BARS scales of general soldiering performance and to the 6 administrative indexes. They felt that there were causal or interactive connections among the factors that made it difficult to assign weights clearly or definitively. The comparatively lower average interjudge correlations among the rankings and ratings of these measures as compared to those obtained for the 10 different types of Project A performance measures could also, of course, reflect disagreement concerning their meaning and importance. For the interested reader, Table 3 presents rank orders of the 10 different types of measures, the 11 BARS scales and the 6 administrative indexes derived from pooling the workshop 2 and 3 data across scaling methods. The rank orders should be interpreted as suggestive of relative importance rather than definitive.

In both the second and third workshops, descriptions of MOS/performance level descriptions were used. These were the same as those used in the first workshop, with one exception. The overall performance scale was changed from one which was behaviorally anchored to one expressed in percentiles (see Exhibit 5). This change was made in recognition of the difficulty of assigning performance-based anchors that would be comparable across MOS in the absence of actual performance data. (Although such performance data will be available in Project A for the 19 sample MOS in early FY 1986, the utility values for the MOS/performance level combinations are scheduled to be obtained in late FY 1985.)

In the first workshop, the officers were asked to rank order 57 MOS/performance level descriptions of enlisted personnel. In the second and third workshops, in addition to rank ordering the described soldiers, the participating officers were asked to assess the relative utility of each of the soldiers in comparison to one particular or standard soldier whose utility was arbitrarily set at 100. The task of the officers was to compare each of the 56 remaining soldiers in turn to the standard soldier and to assign a proportionate utility value to each, given that the standard soldier's value was set at 100. Two standard soldiers were used: the 90th percentile Infantryman (11B) and the 50th percentile Ammunition Specialist (55B). These two MOS/performance level cominbations were, respectively, rank ordered very high and near the median by the first workshop officers. The officers were allowed to assign zero utility values or even negative values if they thought the soldier described would detract from mission accomplishment.

The original intent was to have the workshop 2 and 3 officers do this scaling task twice, first using one standard soldier and then the other. However, time did not permit some of the officers to scale one set of MOS/performance level combinations, let alone two.

Table 3

Cross Workshop/Method Overall Rank
of Performance Measures
(n = 13, Workshops 2 and 3)

| Type of Measure | Overall Rank |
|---|---|
| Job knowledge - Specific | 2.5 |
| Job knowledge - General | 5 |
| Supervisory ratings - Task | 2.5 |
| Peer ratings - Task | 7 |
| Self ratings - Task | 10 |
| Hands-on | 1 |
| Administrative index | 7 |
| Supervisory ratings - General | 4 |
| Peer ratings - General | 7 |
| Self ratings - General | 9 |
| **Supervisory Rating Scale** | |
| Physical fitness | 8 |
| Living and work areas | 11 |
| Controlling behavior | 7 |
| Honesty and integrity | 3.5 |
| Developing skills | 3.5 |
| Leadership | 1.5 |
| Initiative | 1.5 |
| Appearance | 10 |
| Regulations | 6 |
| Maintain equipment | 9 |
| Job knowledge | 5 |
| **201 File Index** | |
| Articles 15 | 1 |
| Training courses | 5.5 |
| Letters/certificates | 3 |
| Reenlistment eligibility | 3 |
| Medals/awards | 5.5 |
| Promotion rate | 3 |

The correlations across the 57 MOS/performance level combinations of the mean rank orders and scale values obtained in both workshops are shown in Table 4. The average interjudge correlations are also shown. These results were considered quite encouraging. The average interjudge correlations and correlations between like utility measures across workshops are sufficiently high to suggest that very reliable average rank and/or ratio scale values could be obtained using about ten judges. The high intercorrelations among the different measures suggest that the final utilty scale values (with appropriate transformations) might be fairly similar across measurement methods.

Table 5 presents the mean scale values assigned the 57 MOS/performance level combinations in workshop 3 by the five officers who used the 50th percentile Ammunition Specialist as the standard. The MOS have been divided in the table into noncombat and combat groups. It is readily apparent that on the average, the combat MOS received higher utilities than the noncombat MOS at all three performance percentiles.

Table 5 also shows for the 19 MOS the differences in average scale values between the 90th and 50th percentile soldiers and the 50th and 10th percentile soldiers. The relationship between performance and utility is generally assumed to be a linear one in the methods used by Schmidt et al. (1979) and Cascio (1982B). Discussions with the officers in the workshops suggested, however, that in combat really good soldiers were worth their weight in gold while really poor soldiers could screw up a whole unit. The data suggest that as performance declines, utility may decline relatively more in the upper percentiles for the noncombat MOS than in the lower percentiles, while for the combat MOS relatively greater decline in utility may take place in the lower percentiles. Obviously such scant data can only be taken as suggestive that nonlinear relationships between performance and utility may exist for some MOS.

In the discussion following the judgment tasks, the officers were questioned concerning the choice between using a high utility MOS/performance level combination (the 90th percentile Infantryman) vs. a median one (the 50th percentile Ammunition Specialist) as the standard for making utility judgments. There was a clear preference for the 90th percentile Infantryman, in part because it was considered easier to scale other MOS between the 0 and 100 points, and in part because Infantryman is the most common and best known Army MOS.

When asked what were the major factors they considered in assigning utilities to the MOS/performance combinations for the wartime scenario given, the officers indicated that potential contribution to unit survival and usefulness in replacing troop losses were foremost in their minds. This is consistent with the ratings given by the Workshop 1 officers of the relative importance of various outcomes (see Table 1).

When asked how general or specific the descriptions of the MOS/performance levels should be, the workshop participants said that most officers think in terms of top, bottom, and mid-level enlisted personnel. That is, either a soldier is good, poor, or somewhere in the middle. They felt that very general performance descriptions would best capture this outlook.

Table 4

Correlations among Types of Utility
Measures for 57 Hypothetical Soldiers
(Workshop 2 and 3 Data)

| | (1) | (2) | (3) | (4) | No. of Judges | Average Interjudge Correlation |
|---|---|---|---|---|---|---|
| Workshop 2 rankings (1) | -- | | | | 6 | .65 |
| Workshop 3 rankings (2) | .93 | -- | | | 7 | .75 |
| Workshop 2 11B - 90% (3) | -.88 | -.87 | -- | | 3 | .61 |
| Workshop 3 11B - 90% (4) | -.87 | -.93 | .79 | -- | 2 | .81 |
| Workshop 3 55B - 50% (5) | -.90 | -.96 | .84 | .85 | 5 | .72 |

Table 5

Scale Values of MOS/Performance Level
Hypothetical Sodliers
(Ammunition Specialist -- 50% = 100; n = 5, Workshop 3

| MOS | Percentile 10 | 50 | 90 | Scale Difference (90-50) | (50-10) |
|---|---|---|---|---|---|
| Administrative Specialist (71L) | 3 | 66 | 98 | 32 | 63 |
| Ammunition Specialist (55B) | 58 | 100 | 153 | 53 | 42 |
| Carpentry & Masonry Specialist (51B) | -8 | 52 | 100 | 48 | 60 |
| Chemical Operations Specialist (54E) | 81 | 150 | 253 | 103 | 69 |
| Food Service Specialist (94B) | 20 | 82 | 140 | 58 | 62 |
| Light Wheel Veh./Power Gen. Mech. (63B) | 29 | 93 | 145 | 52 | 64 |
| Medical Specialist (91B) | 53 | 117 | 197 | 80 | 64 |
| Military Police (95B) | 42 | 95 | 166 | 71 | 53 |
| Motor Transport Operator (64C) | 28 | 94 | 130 | 36 | 66 |
| Petrol. Supply Specialist (76W) | 59 | 94 | 150 | 56 | 35 |
| Radio Teletype Operator (05C) | 44 | 101 | 176 | 75 | 57 |
| TOW/Dragon Repairer (27E) | 48 | 102 | 159 | 57 | 54 |
| Unit Supply Specialist (76Y) | 21 | 71 | 110 | 39 | 50 |
| Util. Heli. Repairer (67N) | 33 | 78 | 130 | 52 | 45 |
| | | | Average | 58 | 45 |
| Infantryman (11B) | 101 | 189 | 260 | 71 | 88 |
| Armor Crewman (19E/K) | 85 | 176 | 250 | 74 | 91 |
| Cannon Crewman (13B) | 88 | 183 | 262 | 79 | 95 |
| Manpads Crewman (16S) | 69 | 174 | 227 | 53 | 105 |
| Combat Engineer (12B) | 79 | 173 | 248 | 75 | 94 |
| | | | Average | 70 | 95 |

The officers felt that the most appropriate judges for weighting measures to arrive at a composite score within an MOS would not be the same judges who would be most appropriate for making cross MOS utility assessments. For weighting measures, they felt that company grade (captains and lieutenants) and senior NCOs from the best units in a variety of commands would make the best judges. For assigning MOS utilities, they felt that field grade officers with recent command experience from various Army Branches would make the best judges. They suggested that officers attending the Army War College and the Command and General Staff School would be good candidates.

## Workshops 4 and 5

The fourth and fifth workshops were conducted for the most part with the field grade officers who had participated in the first and third workshops. Owing to other commitments of some of the earlier participants, complete judgmental data were collected from only two officers who had attended the first workshop and five officers who had attended the third workshop. A sixth officer, who had not attended the third workshop, attended the fifth workshop.

The officers at both workshops were asked to follow a procedure for judging the relative worth or utility of different types of soldiers that we had not tried out before. Using the same wartime scenario and the 57 MOS/performance level combinations used in the third workshop, the officers were asked to judge 228 pairs of MOS/performance level combinations. The judgments were of the form: (_____) soldiers of MOS/performance level combination 1 are equal in overall worth to the Corps in the wartime military situation as (_____) soldiers of MOS/performance level combination 2. The judgmental task was to fill in the two blanks with numbers that would make the two types of soldiers equal in worth. For example, if the two MOS/performance level combinations were 90th percentile Utility Helicopter Repairer (67N) and 50th percentile Combat Engineer (12B), an officer might judge that seven of one type would be worth five of the other. The officers were allowed to put in any number that they liked in order to make the two groups of soldiers equal in worth.

The 228 pairs of MOS/performance level combinations consisted of two types: (1) 57 pairs in which each pair member was from the same MOS but at a different performance level, i.e., 10th, 50th, or 90th percentile (there were 19 MOS x 3 pairs , 10-50, 10-90, and 50-90); and (2) 171 pairs in which each pair member was from a different one of the 19 MOS, with one performance level for each MOS (19 x 18/2 = 171). The 228 pairs were randomized and then presented in the same order to all judges. (Exhibit 6 contains the instructions used and the first page of the judgment record form.)

Scale values for each of the 19 MOS/performance level combinations making up the 171 judgmental pairs were calculated using a ratio scaling procedure described by Torgerson (1958, p. 105-112). This procedure results in a set of scale values whose geometric mean is equal to 1.0. The same procedure was used to separately scale each of the 19 sets of 3 MOS/performance level combinations making up the 57 judgmental pairs. The scale values obtained for the 19 sets of 3 combinations were then transformed to the scale derived from the 171 judgmental pairs in a manner that maintained their original ratio. For example, suppose the 10, 50, and 90th percentile performance levels for MOS A had respective scale values of .5, 1 and 2, when separately scaled as a set of three. Suppose also that the MOS A/10th percentile performance level combination had a scale value of .6 when scaled in the set of 19 (the 171 judgmental pairs). Then the transformed scale values of the three MOS A performance level combinations would be .6, 1.2 and 2.4, respectively.

Scale values for each of the 57 MOS/performance level combinations were first derived from the judgmental data of each officer separately. The geometric mean of the officers' scale values was then calculated for each MOS/performance level combination.[6] Finally, the 57 sale values were divided by the scale value of the 50th percentile Infantryman (11B) in order to make the unit of measurement of the utility scale equal to the value of a 50th percentile Infantryman in the wartime scenario given.

Table 6 presents the average of the officer's scale values obtained for the 57 MOS/performance level combinations using the paired comparison ratio scaling technique described above. Consistent with earlier findings, the combat MOS generally have higher utility ratings at each of the three performance levels (10, 50, and 90th percentile) than the noncombat MOS. However, the difference in utility scale values within an MOS from the 90th to 50th percentile performance level is greater for all 19 MOS than the difference in utility scale values from the 50th to 10th percentile performance level. This is especially true of the combat MOS which on the average showed the highest declines in utility values from the 90th to 50th percentile performance levels. The inconsistency of these results with those cited earlier (see Table 5 from workshop 3) may be more attributable to the scaling method used than to the sample of officers involved, since the officers whose judgments were pooled to arrive at the workshop 5 scale values overlapped considerably with the officers in workshop 3. (That

---

[6] Prior to averaging, the 57 scale values of each officer were multiplied by a constant which made the geometric mean of each officer's scale values equal to 1.0. Averaging the officers' values for each MOS/performance level combination through using the geometric mean maintained the geometric mean of the set of 57 values at 1.0.

Table 6

Scale Values of MOS/Performance Level
Hypothetical Soldiers
(50th Percentile Infantryman = 1.0; n = 8, Workshops 4 and 5

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

there may be considerable variation in relative utility values as a function of scaling method is also suggested by the data from workshops 6 and 7. See Table 10, page 63.)

The average interjudge correlation between the scale values of the 8 officers taken across the 57 combinations was .61. This value, though not as high as that obtained for the scaling methods tried out in workshop 3 (see Table 4), was considered encouraging enough to try out the scaling method again in workshops 6 and 7.

As 5 of the 6 officers in Workshop 5 had rank ordered the 57 MOS/performance level combinations using the same wartime scenario in workshop 3, one and one-half months earlier, it was of interest to determine how reliable their average rankings were. The correlation between first and second average rankings of the 5 officers across the 57 combinations was .98. Another indication of the stability of the average rankings is the average interjudge correlation obtained among the rank orders of the 6 officers. The obtained average, .79, is slightly higher than the average obtained for workshop 3 (.75). Both average interjudge correlations indicate that the average rank ordering based on 10 judges would probably have a reliability of .95 or better.

After the six officers in workshop 5 finished scaling the MOS/performance level combinations, they were asked to rerank the 57 combination cards under a peacetime scenario (see Exhibit 3). The peacetime scenario was set in Europe under current conditions and emphasized maintaining force readiness. Our intent was to determine the impact of scenario differences on the utility values of specific MOS/performance level combinations.

Table 7 shows the MOS/performance level combinations having differences in average assigned rank of 10 or more under the wartime vs. peacetime scenarios. The trend in the data from the six officers is clear--low performance level combat troops are ranked higher in wartime than peacetime, while high performance level support personnel are ranked lower in wartime than peacetime.

The differences in average utility ranks found in Table 7 are certainly not surprising. In fact, if they had been otherwise, one might question the ability of the officers to rank order MOS/performance level combinations in terms of their utility. But they do raise the interesting question of how a computerized selection and assignment procedure can best use utilities if such utilities are a function of the scenario context in which the judgments of utility are made. It may be necessary to use utilities obtained through a number of probable scenarios or to decide upon one particular scenario as the context for the utility judgments. On the other hand, there may not be a significant difference in who gets selected and classified into given MOS using utilities obtained under different scenarios. Not only will other factors, such as the number of applicants, their predictor score distributions and Army personnel requirements and available positions, influence the selection and assignment process, but the average utility rankings themselves may in general be fairly highly correlated even if there are individual MOS/performance level combinations that are ranked quite differently. The correlation across the 57

Table 7

MOS/Performance Level Hypothetical Soldiers with Large Mean Wartime vs. Peacetime Differences in Rank Order (n = 6, Workshop 5)

| | Mean Rank | |
|---|---|---|
| | Wartime | Peacetime |
| **Wartime Higher Than Peacetime** | | |
| Cannon Crewman/10th percentile | 29 | 39 |
| Cannon Crewman/50th percentile | 10 | 20 |
| Chemical Opers. Spec./10th percentile | 35 | 48 |
| Infantryman/10th percentile | 25 | 40 |
| Infantryman/50th percentile | 10 | 20 |
| Armor Crewman/10th percentile | 25 | 37 |
| Manpads Crewman/10th percentile | 31 | 42 |
| **Peacetime Higher Than Wartime** | | |
| Administrative Spec./10th percentile | 56 | 45 |
| Administrative Spec./50th percentile | 46 | 28 |
| Administrative Spec./90th percentile | 36 | 17 |
| Carpentry & Masonry Spec./50th percentile | 50 | 39 |
| Carpentry & Masonry Spec./90th percentile | 41 | 26 |
| Food Service Spec./50th percentile | 41 | 25 |
| Food Service Spec./90th percentile | 30 | 12 |
| Unit Supply Spec./90th percentile | 28 | 14 |

combinations of the average rank assigned by the six officers under the wartime and peacetime scenarios was .85. Computer simulations using different utility values and realistic operational constraints may eventually be needed to determine the practical significance of scenario differences. Some high level policy decisions may also be needed concerning the scenario(s) the computerized personnel system should be geared toward maximizing performance therein.

The workshop 5 participants also rank ordered and scaled eight performance factors as to their importance in forming a composite or overall measure of performance under the wartime and peacetime scenarios. The scaling procedure, which was also used in workshop 3, called for the officers to assign 100 points to the top ranked performance factor, zero points to factors (if any) they thought should not be weighted into the composite, and intermediate points to the remaining factors which reflected their relative importance to the top-rated factor and to each other.

Table 8 gives the average values given the performance factors by the six participating officers. As seen in the table, the average rank order and scale values obtained under the wartime and peacetime scenarios were fairly similar for seven of the eight factors. The factor for which there was a substantial discrepancy under the wartime vs. peacetime scenario was "performance under adverse conditions." As might be expected, this factor was ranked and scaled higher under the wartime scenario. The practical significance of this particular finding may not be great, however. The correlation across the eight factors between the war and peacetime average rankings was .92; the correlation between the war and peacetime average scale values was .93. These high correlations indicate that composite scores derived from applying weights obtained under a wartime scenario to the performance factors would most probably correlate very highly (.95 or above) with composites derived from applying weights obtained under a peacetime scenario.

After they had completed the judgmental tasks, discussions were held with the workshop participants on a number of utility issues. Consistent with the results of later analyses of their response data, the officers reported that their assignments of importance weights to the eight performance factors under the two scenarios were about the same with the exception of the factor, performance under adverse conditions. Also as the data indicated, they reported that they assigned higher ranks to high performance personnel in support MOS under the peacetime than under the wartime scenario.

Some officers reported being concerned when using the paired comparison ratio scaling method that they were being inconsistent in assigning numbers across the judgmental pairs of MOS/performance level combinations. We assured the officers that inconsistency could be expected within that type of judgment series. (The instructions were later modified in workshops 6 and 7 to stress that it was not necessary to strive for consistency in making these kinds of judgments.)

When asked what MOS/performance level soldiers might best be used as a standard or unit in measuring the utility of other soldiers, the officers generally agreed that the 50th percentile Infantryman would be the best

Table 8

Mean Ranks and Scale Values Assigned Performance Factors under Wartime and Peacetime Scenarios (n = 6, Workshop 5)

| Performance Factor | Wartime | | Peacetime | |
|---|---|---|---|---|
| | Rank | Scale | Rank | Scale |
| Dependability in fulfilling assignments | 1.8 | 96.5 | 1.5 | 98.2 |
| Maintenance of equipment and quarters | 4.8 | 69.2 | 4.3 | 75.5 |
| Performance of MOS specific tasks | 2.5 | 93.3 | 2.0 | 95.0 |
| Commitment to Army regulations and traditions | 7.3 | 37.5 | 7.0 | 47.0 |
| Reenlistment eligibility and likelihood | 7.7 | 21.7 | 7.8 | 26.7 |
| Performance of common soldiering tasks | 5.2 | 66.7 | 4.5 | 68.8 |
| Peer support and cooperation | 4.5 | 68.3 | 4.7 | 66.3 |
| Performance under adverse conditions | 2.2 | 92.5 | 4.2 | 69.8 |

choice. They felt that not only are there more Infantrymen than soldiers in any other MOS, but that officers in general had a good understanding of what an average Infantryman was like and what he could do. The officers were also asked what their reaction would be to expressing the differential worth or utility of soldiers in wartime in terms of dollars. Some of the officers reacted very negatively to this concept, citing possible adverse political consequences as well as internal Army morale problems if dollar figures were placed on soldiers' worth. Specifically, they were concerned that the dollar figures might be misinterpreted as the Army's evaluation of the worth of soldiers' lives in wartime. In short, they felt that a soldier's worth to his country could not be evaluated in terms of dollars, especially in wartime.

## Workshops 6 and 7

When the officers in workshops 6 and 7, which were held in Europe, were asked the same question concerning the use of a utility dollar metric their general reaction was also strongly negative. They, like the officers in earlier workshops, agreed that the 50th percentile Infantryman would make the best standard against which the utility of soldiers in other MOS/performance level combinations could be judged. They also agreed that the performance factor whose judged weight was most differentially impacted by wartime vs. peacetime scenarios was performance under adverse conditions. They further agreed that high performance support personnel are more important in peacetime. In wartime, however, they stated that even low performance combat arms personnel are important.

Thirteen officers attended workshops 6 and 7. They were all captains and majors while the earlier workshop participants had all been majors and lieutenant colonels. The consistency of the opinions expressed by the officers in the discussion periods, despite the differences in grade levels and locations, points to a fairly well-shared frame of reference on the part of Army officers.

This common viewpoint was also reflected in the results of the analyses of the workshop data. The workshop 6 and 7 participants were asked to make essentially the same types of judgments made by earlier workshop participants. Only this time the utility of 95 MOS/performance level combinations was judged (5 performance levels--10, 30, 50, 70, and 90th percentile--for each of the 19 MOS) instead of 57 combinations. The correlation across the average paired compairson ratio scale values of the 57 combinations that were common between Workshops 4 and 5 (lieutenant colonels and majors) and Workshops 6 and 7 (captains and majors) was .94. The correlation between the average scale values assigned the eight performance factors (see Table 8) under the wartime scenario by the two groups of officers was .97; under the peacetime scenario, the correlation was .90.

The mean of the rank orders assigned the 95 MOS/performance level combinations under the war and peacetime scenarios by the 13 workshop 6 and 7 officers are shown in Table 9. The MOS in the table have been placed in three groups. The first group contains mostly combat MOS. All the MOS/ performance level combinations involving these MOS had higher average rank

Table 9

Mean Rank Order of MOS/Performance Level
Combinations Under Wartime and Peacetime Scenarios
(n = 13, Workshops 6 and 7)

| | | Performance Percentile | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 70 | 90 |
| **Ranked Higher in Wartime Scenario** | | | | | | |
| Infantryman | W | 64 | 43 | 23 | 13 | 6 |
| (11B) | P | 83 | 69 | 47 | 33 | 17 |
| Armor Crewman | W | 66 | 48 | 25 | 15 | 7 |
| (19E/K) | P | 83 | 67 | 47 | 32 | 16 |
| Cannon Crewman | W | 68 | 48 | 27 | 17 | 9 |
| (13B) | P | 83 | 68 | 51 | 32 | 16 |
| Chemical Operations | W | 73 | 57 | 40 | 24 | 14 |
| Specialist (54E) | P | 86 | 69 | 52 | 39 | 19 |
| Radio Teletype | W | 77 | 60 | 41 | 26 | 14 |
| Operator (05C) | P | 84 | 72 | 53 | 36 | 15 |
| Combat Engineer | W | 72 | 52 | 32 | 27 | 14 |
| (12B) | P | 87 | 68 | 49 | 34 | 19 |
| Manpads Crewman | W | 74 | 53 | 37 | 24 | 15 |
| (16S) | P | 85 | 68 | 54 | 35 | 18 |
| **Ranked Higher in Peacetime Scenario** | | | | | | |
| Administrative Specialist | W | 88 | 77 | 68 | 57 | 41 |
| (71L) | P | 80 | 57 | 40 | 25 | 7 |
| Unit Supply Specialist | W | 82 | 73 | 54 | 42 | 27 |
| (76Y) | P | 76 | 61 | 39 | 22 | 7 |
| Light Wheel Veh./Power | W | 79 | 63 | 48 | 33 | 23 |
| Gen. Mech. (63B) | P | 79 | 61 | 42 | 27 | 10 |
| Food Service Specialist | W | 83 | 70 | 56 | 46 | 35 |
| (94B) | P | 81 | 60 | 44 | 28 | 12 |
| Carpentry & Masonry | W | 91 | 84 | 75 | 67 | 56 |
| Specialist (51B) | P | 84 | 65 | 50 | 34 | 20 |

Table 9 (cont.)

Mean Rank Order of MOS/Performance Level Combinations Under Wartime and Peacetime Scenarios (n = 13, Workshops 6 and 7)

| | | Performance Percentile | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 70 | 90 |
| **Mixed** | | | | | | |
| Medical Specialist | W | 74 | 57 | 45 | 26 | 15 |
| (91B) | P | 82 | 61 | 42 | 23 | 7 |
| TOW/Dragon | W | 80 | 62 | 50 | 35 | 27 |
| Repairer (27E) | P | 83 | 67 | 48 | 33 | 15 |
| Utility Helicopter | W | 76 | 61 | 45 | 34 | 23 |
| Repairer (67N) | P | 81 | 65 | 43 | 28 | 18 |
| Motor Transport Operator | W | 80 | 63 | 52 | 39 | 33 |
| (64C) | P | 82 | 65 | 45 | 29 | 13 |
| Military Police | W | 81 | 66 | 53 | 41 | 28 |
| (95B) | P | 83 | 67 | 45 | 31 | 11 |
| Petrol. Supply Specialist | W | 79 | 64 | 50 | 32 | 20 |
| (76W) | P | 82 | 63 | 48 | 30 | 12 |
| Ammo. Specialist | W | 78 | 65 | 48 | 33 | 22 |
| (55B) | P | 84 | 72 | 55 | 37 | 19 |

orders under the wartime than under the peacetime scenario. All the MOS/performance level combinations in the second group of MOS were ranked higher under the peacetime than wartime scenario. Soldiers in these MOS are generally not expected to be in combat. The average rank orders of the MOS/performance level combinations in the third group of MOS were all higher under peacetime than wartime at the upper levels of performance, but were all lower under peacetime than wartime at the lower levels of performance. Soldiers in these MOS generally have a higher probability of being in a combat situation than soldiers in the second group of MOS.

These data were consistent with the workshop 4 and 5 findings reported earlier (see Table 7) and the statements made during the discussion periods--soldiers at low performance levels that are likely to be involved in combat are assigned relatively higher utility under a wartime scenario, while soldiers at high performance levels that are unlikely to be involved in combat are assigned relatively higher utility under a peacetime scenario. As was discussed earlier, the practical significance of this utility scenario differential in terms of how a computerized personnel system would select and classify large groups of Army applicants could be explored through computer simulations. However, since the correlation across the 95 combinations of the utility values under the two scenarios may be quite high (the correlation of average rank orders was .83 in the workshop 6 and 7 data and .85 for the comparable workshop 4 and 5 data), the simulations may well result in relatively minor scenario-derived differences.

Another factor that may impact the MOS/performance level combination utility values is the scaling procedure used. In workshops 6 and 7, 12 of the officers scaled the 95 combinations in two ways. One method was the paired comparison ratio procedure used by the workshop 4 and 5 participants. They also scaled the 95 combinations using the subjective estimation procedure employed by the Workshop 3 and 4 participants. In this method one combination is given a utility value of 100 and the other combinations are assigned scale values which reflect their respective proportionate utilities. The combination assigned the value of 100 was the 90th percentile Infantryman. The scales obtained by the two methods were then transformed to ones in which the 50th percentile Infantryman had a utility value of 1.0.

Table 10 shows the scale values of the 95 MOS/performance level combinations obtained through using both methods. The scale values obtained from the two methods are quite similar at the lower performance levels. However, with the exception of the Infantryman and Armor crewman MOS, the utility scale values for the higher performance levels obtained from the subjective estimation procedure are higher than those obtained using the paired comparison ratio scaling technique.

Examination of the utilities assigned to the performance levels within MOS revealed that on the average, for both the combat and noncombat MOS the subjective estimation utility values had a somewhat greater decline in the lower half of the performance levels (between the 50th and 10th percentiles) than in the upper half (between the 90th and 50th percentiles). The paired comparison utility values, on the other hand, on the average had a somewhat greater decline in the upper half of the performance levels than in the lower half for both kinds of MOS.

Table 10
Mean Values of MOS/Performance Level Combinations
Using Subjective Estimate and Paired Comparison
Ratio Scaling Techniques
(n = 12, workshops 6 and 7)

| MOS | | Performance Percentile 10 | 30 | 50 | 70 | 90 |
|---|---|---|---|---|---|---|
| Administrative Specialist (71L) | SE | -.07 | .29 | .47 | .74 | .86 |
| | PC | .09 | .16 | .24 | .31 | .45 |
| Ammunition Specialist (55B) | SE | .12 | .46 | .69 | .90 | 1.13 |
| | PC | .12 | .26 | .38 | .52 | .73 |
| Cannon Creman (13B) | SE | .30 | .69 | .93 | 1.24 | 1.49 |
| | PC | .24 | .41 | .64 | .90 | 1.28 |
| Carpentry & Masonry Specialist (51B) | SE | .00 | .09 | .37 | .61 | .80 |
| | PC | .07 | .11 | .18 | .24 | .38 |
| Chemical Operations Specialist (54E) | SE | .20 | .53 | .86 | 1.16 | 1.38 |
| | PC | .16 | .35 | .46 | .67 | .96 |
| Combat Engineer (12B) | SE | .20 | .65 | .96 | 1.22 | 1.52 |
| | PC | .19 | .38 | .57 | .77 | 1.05 |
| Food Service Specialist (94B) | SE | .09 | .33 | .59 | .83 | 1.04 |
| | PC | .11 | .16 | .27 | .38 | .50 |
| Infantryman (11B) | SE | .29 | .71 | 1.00 | 1.30 | 1.58 |
| | PC | .39 | .63 | 1.00 | 1.53 | 2.16 |
| Light wheel Veh./Power Gen. Mech. (63B) | SE | .17 | .51 | .66 | 1.02 | 1.24 |
| | PC | .13 | .24 | .37 | .50 | .65 |
| Armor Crewman (19E/K) | SE | .40 | .68 | 1.03 | 1.26 | 1.60 |
| | PC | .25 | .48 | .73 | 1.14 | 1.63 |
| Manpads Crewman (16S) | SE | .19 | .57 | .83 | 1.09 | 1.38 |
| | PC | .16 | .31 | .45 | .65 | .96 |
| Medical Specialist (91B) | SE | .17 | .48 | .79 | 1.07 | 1.37 |
| | PC | .15 | .30 | .42 | .62 | .95 |
| Military Police (95B) | SE | .15 | .47 | .71 | .97 | 1.20 |
| | PC | .16 | .26 | .36 | .52 | .74 |
| Motor Trans. Operator (64C) | SE | .06 | .39 | .59 | .83 | .97 |
| | PC | .13 | .21 | .33 | .43 | .65 |
| Petrol. Supply Specialist (76W) | SE | .16 | .51 | .72 | .82 | 1.11 |
| | PC | .13 | .25 | .39 | .52 | .78 |
| Radio Teletype Operator (05C) | SE | .13 | .54 | .77 | 1.09 | 1.30 |
| | PC | .16 | .26 | .42 | .53 | .80 |
| Tow/Dragon Rep. (27E) | SE | .10 | .53 | .74 | .99 | 1.33 |
| | PC | .16 | .28 | .43 | .56 | .78 |
| Unit Supply Specialist (76Y) | SE | .08 | .40 | .60 | .91 | 1.07 |
| | PC | .12 | .22 | .34 | .50 | .69 |
| Utility Helicopter Repairer (67N) | SE | .15 | .49 | .82 | 1.06 | 1.32 |
| | PC | .17 | .30 | .43 | .62 | .90 |

SE: Slightly greater decline in lower half than in upper for both combat and noncombat.

PC: Slightly greater decline in upper half than lower half for both combat and noncombat but somewhat larger for combat.

As in the case of the scenario differences, these scaling method differences may or may not have practical significance. The correlation between the mean values assigned the 95 combinations by the two methods was .91. If further research indicates that the scaling method differences found here hold across different groups of officers and MOS/performance level combinations, then it would be advisable to try to assess through computer simulations the likely impact of the scale differences obtained on the selection and classification process.

An examination was made of the amount of agreement among the judges in the ranks and scale values assigned the 95 MOS/performance level combinations under the wartime scenario. Comparing the standard deviations of the values assigned by the officers to each of the combinations led to the identification of the combinations that were generally ranked and scaled most similarly by the officers and those that they disagreed the most about. Table 11 shows these combinations. It is of interest to note that in general the highest disagreement in assigning scale values occurred with high performance level noncombat MOS combinations, whereas the highest agreement in assigning scale values occurred with low performance level noncombat MOS combinations. The perceived usefulness in a wartime scenario of high performance noncombat MOS personnel apparently varies considerably among officers.

In general, however, as noted earlier, the Army officers seem to have a fairly common frame of reference. The median intercorrelations among the officers for the wartime rank orders and scaling values ranged from .76 to .80. Average scale values based upon the judgments of 10 or more officers should therefore have reliabilities of .95 or higher.

## Conclusions

We began these series of exploratory utility workshops not knowing whether Army officers would be willing and able to assign differential utility values across MOS and performance levels. Perhaps our most significant finding is that they are! Perhaps our next most significant finding is that the utility values assigned by different officers are sufficiently alike to indicate that fairly stable scale values could be obtained from averaging across officers' judgmental data. Taken together. these two findings point to the feasibility of setting differential MOS/performance level utility values that could be used to help guide a computerized enlisted personnel selection and classification system.

This exploratory research revealed, however, several problems that need addressing before such an outcome can be achieved. For one, the utility values that are assigned by the officers vary as a function of the scenario or context in which the evaluative judgments are made. For another, the utility values assigned by the officers may vary as a function of the judgment or scaling procedure used to obtain the values. Moreover, the officers tend to agree more about the utility of certain MOS/performance level soldiers in a wartime setting than others. Why? Is the role of certain types of soldiers in wartime not as well defined or clear as the role of other soldiers? Should these roles be clarified before final utility values are obtained?

There are other as yet unexplored questions. Will the utility value assigned a given MOS/performance level combination vary as a function of the other combinations with which it is being judged? Do officers with certain backgrounds or from certain Army career branches systematically differ in their utility assessments? Would group judgmental processes, e.g., the Delphi method, yield better utility values than the individually based methods employed in this exploratory research? Would it be better to scale MOS/performance level combinations and performance factors separately as we have done in our research to date; or would it be better to scale both at the same time through application of a conjoint scaling technique?

As with most exploratory research we have perhaps raised more questions than we have answered. But we have learned some things as well (see exhibit 7 for a list of 10 items learned). The scenario(s) used should be free of detail which suggest greater or less utility for certain specific MOS. Utilities of soldiers in wartime should probably not be expressed in terms of dollars; an apparently acceptable metric would be the utility of a 50th percentile Infantryman (his value for the survival of the unit and in replacing troop losses is much more readily apparent). Directions to the judges should be reassuring concerning inconsistencies that can possibly occur in a long series of judgments. Certain performance factors are consistently given higher weights than others in forming a composite criterion. However, scenario differences in how various performance factors are weighted to arrive at a composite score are probably not a major concern.

As discussed earlier in this paper, some of the problems we have identified, e.g., scenario effects, may have little practical significance in terms of how a computerized enlisted personnel selection and classification system would process Army applicants under operational constraints. As further research in establishing utility values for all Army MOS is pursued, we hope to examine through sensitivity analyses and computer simulations how differences in the utilities of MOS/performance level combinations affect system output. These results should help us to further focus our research.

## REFERENCES


Brogden, H. E. (1949). When testing pays off. Personnel Psychology. 2, 171-183.

Cascio, W. F. (1982a). Applied psychology in personnel management (2nd Ed.) Reston, VA: Reston Publishing.

Cascio, Wayne F. (1982b). Costing human resources: The financial impact of behavior in organizations. Boston: Kent Publishing.

Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions (Rev. ed.) Urbana: University of Illinois Press.

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

Roulon, P. J., Tiedeman, D. V., Tatsuoka, M. M., & Langmuir, C. R. (1967). Multivariate statistics for personnel classification. New York: John Wiley & Sons, Inc.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work force productivity. Journal of Applied Psychology, 64, 609-626.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. Journal of Applied Psychology, 23, 565-578.

Torgerson, Warren S. (1958). Theory and methods of scaling. New York: John Wiley & Sons, Inc.

Ward, J. H., Haney, D. L., Hendrix, W. H., & Pina, M. (1978). Assignment procedure in the Air Force procurement management information system (AFHRL-TR-78-30). Brooks AFB, TX: Air Force Human Resources Laboratory.

EXHIBIT 1

DIRECTIONS FOR RANK ORDERING THE UTILITY
OF DIFFERENT ENLISTED PERSONNEL

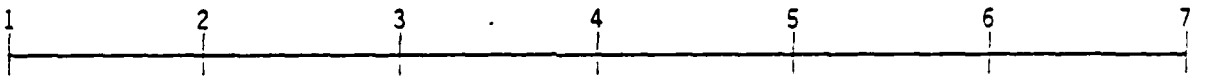
You will be given a set of 57 cards. On each card there is a short description of a different soldier. The descriptions contain the following information about each soldier:

(1) A brief summary of the soldier's MOS and associated job duties; and

(2) An overall rating of the effectiveness of the soldier.

The summary MOS statements were taken from AR 611-201, which gives descriptions of all Army MOS. The overall effectiveness ratings were taken from newly developed scales based upon behavioral incidents of effective/ineffective enlisted personnel performance. The overall effectiveness rating scale looks like this:

Overall Soldier Effectiveness:




| BELOW STANDARD | ADEQUATE/MID-RANGE | SUPERIOR |
|---|---|---|
| Performs poorly in important effectiveness areas; does not meet standards and expectations for adequate soldier performance. | Adequately performs in important effectiveness areas; meets standards and expectations for adequate soldier performance. | Performs excellently in all or almost all effectiveness areas; exceeds standards and expectations for soldier performance. |

The soldiers described on the 57 cards all received accurate overall effectiveness ratings of either "2", "4", or "6" on the above scale. The soldiers are from 19 MOS and there are three soldiers from each MOS (3x19 = 57). Each of the three soldiers within an MOS received a different overall effectiveness rating.

EXHIBIT 1 (cont.)

The cards have been placed in random order. Your task is to rank order the cards so that the soldier who has the most overall utility for the Army is ranked first, the soldier who has the second most overall utility for the Army is ranked second, etc. Please follow these directions in making your rankings:

(1) Briefly look through the set of cards to get an understanding of the types of jobs the soldiers do. The list of sample MOS can also be used for this purpose.

(2) Then sort the soldiers into 7 piles with about 6-10 soldiers in each pile. The first pile should contain the soldiers that you think have the most overall utility, the last pile those with the least utility to the Army, and the in-between piles those with intermediate utility.

(3) Rank order the soldiers in the first pile, placing the one with the most utility for the Army first.

(4) Then rank order the soldiers in the last pile, placing the soldier with the least utility last.

(5) Then rank order the soldiers in piles 2, 6, 3, 5, and 4 in that order.

(6) Put the piles together into one stack in the order 1, 2, 3, 4, 5, 6, 7 and go through the cards one at a time from the soldier with the most utility to the soldier with the least, making any changes in the rank order that you feel are appropriate.

(7) Then go through the set of cards in reverse order, that is, from the soldier with the least utility to the one with the most, again making any changes in the rank order that you feel are appropriate.

(8) When you are satisfied with your rank ordering, please place a rubberband securely around the cards to preserve the order.

Thank you for your cooperation.

Sample Card

MEDICAL SPECIALIST

SUMMARY: Supervises dispensary or field medical facilities, administers emergency medical treatment to battlefield casualties, assists with inpatient and outpatient care and treatment, and assists with technical and administrative management of medical treatment facilities.

DUTIES: Performs routine field medical activities and patient care procedures.

OVERALL EFFECTIVENESS: Adequate/Mid-Range (Rating: 4)

EXHIBIT 2

## SAMPLE MILITARY OCCUPATIONAL SPECIALTIES

ADMINISTRATIVE SPECIALIST - 71L

Summary: Supervises or performs administrative, clerical and typing duties.

Duties: Performs typing, clerical and administrative duties.

AMMUNITION SPECIALIST - 55B

Summary: Supervises, performs, or assists in ammunition storage, receipt issue, stock control, accounting, and maintenance operations.

Duties: Assists in receipt, storage, issue, and maintenance of ammunition, ammunition components, and explosives.

CANNON CREWMAN - 13B

Summary: Supervises or serves as member of field artillery cannon unit.

Duties: Participates in emplacement, laying, firing, and displacement of field artillery cannons.

CARPENTRY AND MASONRY SPECIALIST - 51B

Summary: Performs general and heavy carpentry and masonry duties in fabrication, erection, maintenance, and repair of wooden and masonry structures, and variety of wooden articles.

Duties: Performs basic carpentry and masonry duties associated with construction activities.

CHEMICAL OPERATIONS SPECIALIST - 54E

Summary: Operates decontamination equipment and supervises operation of decontamination, smoke and flame equipment; assists in establishment, administration, and application of nuclear, biological, and chemical defense measures and offensive chemical and nuclear operations.

Duties: Decontamination military equipment, material, supplies, and terrain.

COMBAT ENGINEER - 12B

Summary: Commands, serves, or assists as member of team, squad, section, or platoon engaged in providing combat engineering support to combat forces.

Duties: Assists combat engineers by performing combat construction, combat demolitions, and related duties.

FOOD SERVICE SPECIALIST - 94B

Summary: Supervises or prepares and cooks food in field, garrison, or central food preparation activities.

Duties: Prepares and cooks food.

INFANTRYMAN - 11B

Summary: Leads, supervises and serves as member of an infantry activity employing individual weapons and machineguns in offensive and defensive combat operations.

Duties: Closes with and destroys enemy personnel weapons and equipment.

LIGHT WHEEL VEHICLE/POWER GENERATOR MECHANIC - 63B

Summary: Performs and supervises organizational maintenance and recovery operations on light wheel vehicles (prime movers designated as five ton or less and their associated trailers), tactical power generation equipment, and associated items. Supervises organizational maintenance and recovery operations on track and heavy wheel vehicles and materials handling equipment (MHE).

Duties: Troubleshoots and performs organizational maintenance or internal combusti

M1 ARMOR CREWMAN - 19E

Summary: Leads, supervises, or serves as member of M1 armor unit in offensive and defensive combat operations.

Duties: Loads and fires tank main gun.

MANPADS CREWMAN - 16S

Summary: Supervises or serves as member of MANPADS (Man Portable Air Defense System) missile unit.

Duites: Prepares and fires MANPADS missile.

MEDICAL SPECIALIST - 91B

Summary: Supervises dispensary or field medical facilities, administers emergency medical treatment to battlefield casualties, assists with inpatient and outpatient care and treatment, and assists with technical and administrative management of medical treatment facilities.

Duties: Performs routine field medical activities and patient care procedures.

MILITARY POLICE - 95B

Summary: Supervises or provides law enforcement activities, preserves military control, provides security, controls traffic, quells disturbances, protects property and personnel, handles prisoners of war, refugees, or evacuees and investigates incidents.

Duties: Enforces traffic regulations and law and order, exercises military control and discipline and guards prisoners of war.

MOTOR TRANSPORT OPERATOR - 64C

Summary: Supervises or operates wheel vehicles to transport personnel and cargo.

Duties: Operates wheel vehicles.

PETROLEUM SUPPLY SPECIALIST - 76W

Summary: Supervises or receives, stores, accounts and cares for, dispenses, issues, and ships bulk or packaged petroleum, oils, and lubricants (POL) products.

Duties: Receives, stores, accounts, and cares for, dispenses, issues, and ships bulk and packaged POL supplies.

RADIO TELETYPE OPERATOR - 05C

Summary: Supervises or operates and installs radio teletypewriter and tape relay equipment in radio teletypewriter and tape relay tactical or administrative communications nets.

Duties: Operates radio teletype equipment to transmit and receive messages.

TOW/DRAGON REPAIRER - 27E

Summary: Supervises or performs direct support and general support maintenance on the TOW and DRAGON missile systems.

Duties: Performs support level maintenance on the TOW and DRAGON missile systems, trainers, nightsights, battery chargers, and system peculiar test and checkout equipment.

UNIT SUPPLY SPECIALIST - 76Y

Summary: Supervises or performs duties involving request, receipt, storage issue, accounting for, and preservation of individual, organizational, installation, and expendable supplies and equipment.

Duties: Receives, stores, issues, accounts for, and preserves supplies in unit.

UTILITY HELICOPTER REPAIRER -67N

Summary: Supervises, inspects, or performs maintenance on utility helicopters, excluding repair of systems components.

Duties: Assists in organizational, direct, and general support (aviation unit, intermediate and depot) maintenance of utility helicopters, excluding repair of system components.

# SCENARIOS USED IN EXPLORATORY UTILITY WORKSHOPS

## WARTIME SCENARIO: FIRST AND SECOND WORKSHOP

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

## WARTIME SCENARIO: THIRD - SEVENTH WORKSHOP

Hostilities have broken out in Europe and your Corps' combat units are engaged. Your Corps' mission is to defend, then re-establish, the host country's border. Pockets of enemy airborne/helliborne and guerilla elements are operating throughout the Corps sector area. Limited initial and reactive chemical strikes have been employed but nuclear strikes have not been initiated. Air parity does exist.

## PEACETIME SCENARIO: FOURTH - SEVENTH WORKSHOP

Europe is in the peacetime condition currently prevailing there. Your Corps mission is to defend and maintain the host country's border should war break out. The potential enemy approximates a combined arms army and has nuclear and chemical capability. Air parity does exist. The Corps has personnel and equipment sufficient to make its mission capable for training and evaluation. The training cycle includes periodic field exercises, command and maintenance inspections, ARTEP evaluations, and individual soldier training/SQT testing.

EXHIBIT 4

## Directions for Assigning Weights to Component Measures to Arrive at Total Scores

A number of different kinds of performance instruments are being developed by Project A to measure the effectiveness of first-tour enlisted personnel. The principal types of measures that will be administered are:

1. Job knowledge test (specific) - Score on a test of the specific items of information required to perform 30 tasks selected for their importance and representativeness of their MOS jobs.

2. Job knowledge test (general) - Score on a test of knowledge elements required to perform MOS tasks in general.

3. Supervisory ratings of performance of major MOS task areas - Soldiers are rated by their supervisors on their performance on major areas of their jobs, e.g., Administrative Specialist (71L) are rated on keeping records; preparing, typing, and proofreading documents; safeguarding classified documents; etc. Score: Average of 2 supervisor ratings.

4. Peer ratings of performance of major MOS task areas - Soldiers are rated by their peers on their performance of major areas of their jobs [see (3) above]. Score: Average of 3 peer ratings.

5. Self ratings of performance of major MOS task areas - Soldiers rate their own performance of major areas of their jobs [see (3) above].

6. Hands-on performance measures - Soldiers are scored on their performance on 15 specific tasks selected for their importance and representativeness of their MOS jobs.

7. Administrative index - Number of awards, letters/certificates of commendation, Article 15/Flag actions, promotion rate, military training courses, and reenlistment eligibility taken from the 201 files maintained for the soldier.

8. Supervisory ratings of general soldiering - Ratings by their supervisors on such factors as leadership, initiative, maintaining equipment and living/work areas, following regulations/orders, etc. Score: Average of 2 supervisor ratings.

9. Peer ratings of general soldiering - Ratings by their peers on the general soldiering factors [see (8) above]. Score: Average of 3 peer ratings.

10. Self ratings of general soldiering - Soldiers rate their own performance on the general soldiering factors [see (8) above].

A total score will be derived for each soldier from the separate scores obtained from each of these measures. These total scores will be our best estimate of the overall effectiveness of the troops whose performance will be measured. We need the assistance of experienced Army officers in determining first, how much weight should be given each type of measure in arriving at the total effectiveness scores; and second, how much weight should be given each part of the separate components of the instruments to arrive at a score for that instrument.

Today we would like to get your judgments about the weights or number of points that component scores should receive in the total. The procedure for assigning these points is as follows:

1) Rank order in terms of importance the components, assigning a "1" to the most important, a "2" to the next most important, etc., until all components have been ranked.

2) After you have recorded the rank orders, assign 100 points to the component that you ranked as most important. Then ask yourself, "If I'm assigning 100 points to this measure, how many points should I assign to the next most important measure." If, for example, you thought that the second measure should receive half the weight of the first, you would assign it 50 points. Continue assigning points in this manner until all components have been weighted.

3) In assigning the points, please keep in mind that the points represent how many times more (or less) important one component is than the others. For example, if you assign 20 points to one component and 5 points to another, that means that you believe that the 20-point component should receive 4 times the weight in the total score as the 5-point component.

4) If you believe a particular component measure should not be used at all in arriving at the total score you should assign it zero points.

5) When you are finished assigning the points, please make sure that they are in the "right" ratio to one another. That is, that the points assigned to all components are in correct proportion to one another.

First, you are being asked to assign weights to the types of measures listed earlier. Then you will be asked to rate the separate scales or components of two of the measures: the supervisory rating scales and the administrative indexes taken from the 201 file. Please follow the same procedure for each set of ratings, first rank ordering the components in terms of the weight you believe they should receive in the total score and then assigning 100 points to the most important component. Points should then be assigned the other components in a manner that reflects their relative importance.

Thanks for your cooperation.

EXHIBIT 4 (cont.)

Name _____ Date _____

Form for Weighting Project A Performance Measures

| Performance Measure | Rank Order | Assigned Value |
|---|---|---|
| Job knowledge (specific) | ( ) | ( ) |
| Job knowledge (general) | ( ) | ( ) |
| Supervisory Ratings of major MOS Task Areas | ( ) | ( ) |
| Peer Ratings of major MOS task areas | ( ) | ( ) |
| Self Ratings of major MOS task areas | ( ) | ( ) |
| Hands-on performance | ( ) | ( ) |
| Administrative index | ( ) | ( ) |
| Supervisory Ratings of general soldiering | ( ) | ( ) |
| Peer ratings of general soldiering | ( ) | ( ) |
| Self ratings of general soldiering | ( ) | ( ) |

EXHIBIT 4 (cont.)

Name _____ Date _____

Form for Weighting Supervisory Ratings on General Soldiering Scales

| General Soldiering Scale | Rank Order | Assigned Value |
|---|---|---|
| Maintaining military standards of physical fitness | ( ) | ( ) |
| Maintaining living and work areas to Army/Unit standards | ( ) | ( ) |
| Controlling own behavior related to personal finances, drugs/alcohol, and agressive acts | ( ) | ( ) |
| Displaying honesty and integrity in job-related and in personal matters | ( ) | ( ) |
| Developing own job and soldiering skills | ( ) | ( ) |
| Performing in leader role, as required, and providing support for fellow unit members | ( ) | ( ) |
| Showing initiative and extra effort on the job/mission/assignment | ( ) | ( ) |
| Maintaining proper military appearance | ( ) | ( ) |
| Adhering to regulations, orders, and SOP, and displaying respect for authority | ( ) | ( ) |
| Checking on and maintaining own weapons/ vehicles/other equipment | ( ) | ( ) |
| Displaying job and soldiering knowledge/skill | ( ) | ( ) |

Name _____ Date _____

Form for Weighting Administrative Indexes
(Data source: 201 File)

| Administrative Index | Rank Order | Assigned Value |
|---|---|---|
| Number of Article 15/Flag actions* | ( ) | ( ) |
| Number of military training courses | ( ) | ( ) |
| Number of letters/certificates of commendation | ( ) | ( ) |
| Reenlistment eligibility | ( ) | ( ) |
| Number of medals/awards | ( ) | ( ) |
| Promotion rate (grades advanced per year) | ( ) | ( ) |

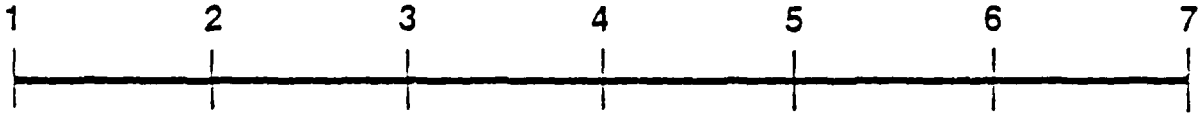
* This index will be reversed so that absence of disciplinary actions will be positively weighted.

EXHIBIT 5

# DESCRIPTION OF SOLDIER PERFORMANCE

## INITIAL

### Overall Soldier Effectiveness:




**BELOW STANDARD**

Performs poorly in important effectiveness areas; does not meet standards and expectations for adequate soldier performance.

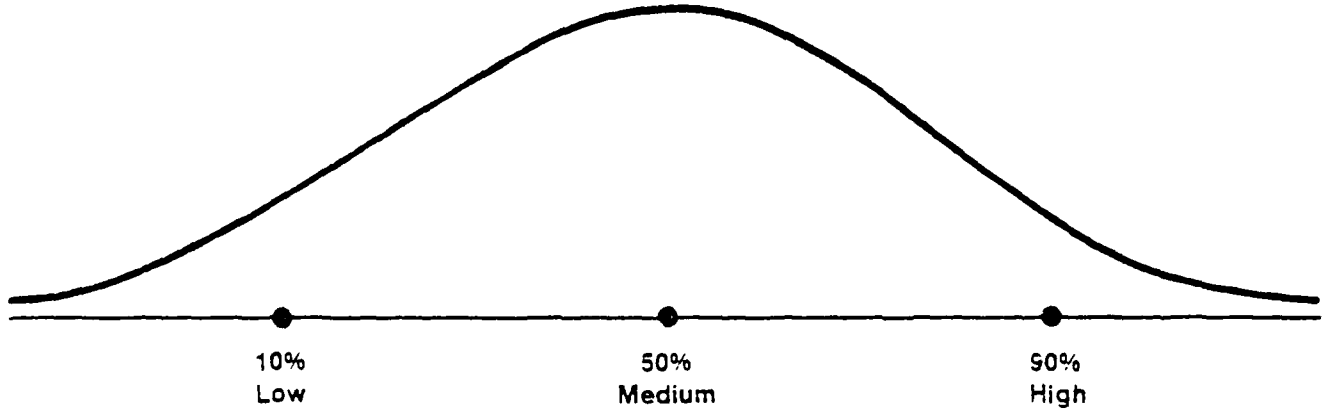
**ADEQUATE/MID-RANGE**

Adequately performs in important effectiveness areas; meets standards and expectations for adequate soldier performance.

**SUPERIOR**

Performs excellently in all or almost all effectiveness areas; exceeds standards and expectations for soldier performance.

## CURRENT




Overall Performances in MOS

...icates low overall performance and 90th percentile indicates high performance.

EXHIBIT 5

## DIRECTIONS FOR JUDGING THE RELATIVE WORTH OF DIFFERENT TYPES OF SOLDIERS

In this procedure you will judge the worth of different types of soldiers in comparison to other types. The procedure essentially involves judging how many soldiers of one type would have the same overall worth as a given number of another type of soldier. The military context for the comparative judgments is the same as before--the outbreak of hostilities in Europe.

The judgments involve the same types of soldier on the cards you have just rank ordered. The judgments are in the form: (_____) soldiers of type X are equal in overall worth to the Corps in the military situation as (_____) soldiers of type Y. Your task is to supply the two missing numbers. For example, suppose the statement reads, "_____ 50tn percentile Military Police - 95B would be equal in worth to _____ 10th percentile TOW/Dragon Repairer - 27E." Let us say that you feel one 50th percentile Military Police would be worth two 10th percentile TOW/Dragon Repairer, then you would put a "1" and a "2" in the two blanks, respectively. Or suppose the statement reads, "_____ 90th percentile Utility Helicopter Repairer - 67N would be equal in worth to _____ 50th percentile Combat Engineer - 12B". If you feel that seven 90th percentile Utility Helicopter Repairer - 67N would be worth five 50th percentile Combat Engineers, then you would put a "7" and a "5" in the two blanks, respectively.

You can put any numbers you like in the two blanks in order to make the two types of soldiers equal in worth. Another example: If you feel that the one type of soldier is worth just a little bit more than another then you could put down the number 100 for the soldier with the slight edge and 101 for

the other. Remember that the number given to the type of soldier that you feel is worth more will always be less than the number given to the type of soldier you feel is not as useful in the military context. Of course, if you feel the two types of soldiers are equal in worth, then you can give each the same number, i.e., a value of "1".

One way of looking at the judgment task is to imagine that the Corps is understrength in all categories of soldiers and that your job is to send equivalent groups of soldiers to the various Corps units. In making your judgments please assume that the Corps could use all the soldiers to the best of their ability. Please also assume that the Corps has sufficient equipment so that the soldiers can be immediately useful in their MOS or as otherwise assigned. Please further assume that the problems of transporting the soldiers and absorbing them in their new units are negligible.

When you have finished making your judgments, please go over them once more and change any numbers that you feel are out of line.

Thank you again for your cooperation.

# EXHIBIT 5 (cont.)

Name_____ Date_____

Judgment Record Form

| | |
|---|---|
| 1. _____90th Percentile Military Police-95E | =_____10th Percentile Military Police-95E |
| 2. _____90th Percentile Motor Transport Operator-64C | =_____50th Percentile Motor Transport Operator-64C |
| 3. _____50th Percentile Chemical Operations Specialist-54E | =_____90th Percentile Combat Engineer-12B |
| 4. _____10th Percentile Military Police-95B | =_____90th Percentile Cannon Crewman-13B |
| 5. _____10th Percentile Petroleum Supply Specialist-76W | =_____50th Percentile Petroleum Supply Specialist-76W |
| 6. _____90th Percentile Petroleum Supply Specialist-76W | =_____10th Percentile Administrative Specialist-71L |
| 7. _____10th Percentile Radio Teletype Operator-05C | =_____10th Percentile M1 Armor Crewman-19E |
| 8. _____50th Percentile Motor Transport Operator-64C | =_____10th Percentile M1 Armor Crewman-19E |
| 9. _____10th Percentile Radio Teletype Operator-05C | =_____90th Percentile Combat Engineer-12B |
| 10. _____50th Percentile Utility Helicopter Repairman-67N | =_____90th Percentile Cannon Crewman-13B |
| 11. _____50th Percentile Ammunition Specialist-55B | =_____10th Percentile Ammunition Specialist-55B |
| 12. _____90th Percentile Cannon Crewman-13B | =_____90th Percentile Medical Specialist-91B |
| 13. _____50th Percentile Motor Transport Operator-64C | =_____50th Percentile Ammunition Specialist-55B |
| 14. _____10th Percentile Food Service Specialist-94B | =_____90th Percentile Medical Specialist-91B |
| 15. _____50th Percentile Ammunition Specialist-55B | =_____50th Percentile Utility Helicopter Repairman-67N |
| 16. _____50th Percentile Chemical Operations Specialist-54E | =_____50th Percentile Motor Transport Operator-64C |
| 17. _____90th Percentile Cannon Crewman-13B | =_____10th Percentile M1 Armor Crewman-19E |
| 18. _____10th Percentile Radio Teletype Operator-05C | =_____50th Percentile Ammunition Specialist-55B |
| 19. _____50th Percentile Motor Transport Operator-64C | =_____90th Percentile Medical Specialist-91B |
| 20. _____10th Percentile M1 Armor Crewman-19E | =_____10th Percentile Carpentry and Masonry Specialist-51B |
| 21. _____90th Percentile Vehicle Power Generator Mechanic-63B | =_____10th Percentile Radio Teletype Operator-05C |
| 22. _____50th Percentile TOW Dragon Repairman-27E | =_____90th Percentile TOW/Dragon Repairman-27E |
| 23. _____90th Percentile Petroleum Supply Specialist-76A | =_____10th Percentile Radio Teletype Operator-05C |
| 24. _____50th Percentile Acquisition Specialist-96B | =_____50th Percentile TOW Dragon Repairman-27E |
| 25. _____90th Percentile Medical Specialist-91B | =_____10th Percentile Military Police-95B |

Table 6

Scale Values of MOS/Performance Level
Hypothetical Soldiers
(50th Percentile Infantryman = 1.0; n = 8, Workshops 4 and 5

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

Table 11

Interjudge Agreement in Ranking and Scaling
MOS/Performance Level Combinations in Wartime
(n = 12, Workshops 6 and 7)

| Highest Agreement* | Lowest Agreement |
|---|---|
| ° Administrative Specialist - 30% | ° Administrative Specialist - 90% |
| ° Carpentry and Masonry Specialist - 10% | ° Carpentry & Masonry Specialist - 90% |
| | ° Chemical Operations Specialist - 90%* |
| ° Food Service Specialist - 10% | ° Food Service Specialist - 90% |
| ° Food Service Specialist - 30% | ° Military Police - 90% |
| ° Light Wheel Vehicle/Power Gen. Mech. - 10% | ° Motor Transport Operator - 90% |
| | ° TOW/Dragon Repairer - 70%* |
| ° Petrol. Supply Specialist - 30% | ° TOW/Dragon Repairer - 90%* |
| ° TOW/Dragon Repairer - 10% * | ° Unit Supply Specialist - 90% |
| ° Unit Supply Specialist - 10% | ° Utility Helicopter Repairer - 90% |

*All but Chemical Operations Specialist and TOW/Dragon Repairer ranked higher in Peacetime than Wartime.

# PERFORMANCE RATINGS AS CRITERIA: WHAT IS BEING MEASURED?

Walter C. Borman
Personnel Decisions Research Institute

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

Elaine D. Pulakos
Personnel Decisions Research Institute

August 1985

## Author Notes

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide information and procedures required to meet military manpower challenges of the future by enabling the Army to enlist, allocate, and retain the most qualified soldiers. This research is funded primarily by Army Project Number 2O263731A791 and is conducted under the direction of the U. S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U. S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

Portions of this paper were presented at the 93rd annual meeting of the American Psychological Association, Los Angeles, California, August, 1985. All statements expressed in this paper are those of the authors and do not represent the official opinions or policies of the U. S. Army Research Institute or the Department of the Army.

The authors wish to thank Betty Shelly for her assistance in the preparation of this manuscript.

Performance Ratings as Criteria: What is Being Measured?

A large Army project is underway to validate new and current predictors of first term soldier performance. A primary goal of this effort is to increase Army organizational effectiveness by improving the soldier job match. This goal will be achieved by developing a comprehensive set of selection and classification measures (predictors) and performance criteria and then empirically demonstrating relationships between these predictor and criterion measures. The need to define and obtain reliable and valid measures of performance is clearly essential to this effort.

The principal methods of performance measurement being developed for this project are (a) hands-on, task proficiency tests (b) job knowledge tests, and (c) supervisor and peer ratings of performance. Later in the project, the multiple measures of performance will be combined into a single composite or composites to measure a soldier's effectiveness on the job.

The focus of this paper is on the "meaning" of peer and supervisory ratings as measures of job performance. Specifically, what these ratings are measuring and relationships between ratings and other kinds of criterion measures. Presently, relatively little research is available to address these issues. One exception is a recent paper by Hunter (1983). In a meta-analysis of 14 studies, Hunter used causal analysis techniques to identify relationships among four variables relevant to work performance: cognitive ability, job knowledge, and job performance as measured by job sample tests and by supervisor ratings. The analysis showed that supervisor ratings were related to both task proficiency and job knowledge required for effective performance, but these relationships were quite low. In the model, the multiple correlation for the prediction of supervisor overall job effective-

ness ratings was 0.42, even after error of measurement was removed. Thus, factors other than those examined by Hunter would appear to account for a large portion of the variance in ratings. An evaluation of the usefulness or "meaning" of job performance ratings as criteria requires information on what these "other influences" are (Guion, 1983).

Several factors have been proposed as having potential to influence job performance ratings. Broadly speaking, these include characteristics of the rater and ratee, the context in which the appraisal is conducted, and various rater/ratee interpersonal relationship factors (Landy & Farr, 1980; Ilgen & Feldman, 1983). While investigations have been conducted to examine the influence of a number of these factors in performance ratings, relatively little research has focused on the potential effects of rater/ratee interpersonal factors on performance evaluations.

In the present research, items pertinent to relationship factors (e.g., friendship/liking) along with ratee personal characteristics (e.g., social skill) were included on the supervisor and peer ratings instruments. This part of the work can be viewed as employing a policy capturing framework (Zedeck & Kafrey, 1977) in that relationships between ratings on these items/dimensions and job performance ratings will provide clues about the relative influence of these factors on performance judgments. In addition, basic demographic data (e.g., race) and job history information (e.g., months in unit) were collected. Correlations between these measures and job performance ratings were also examined in this research.

In summary, the purpose of the research was two fold: a) to evaluate relationships between job performance ratings and other measures of job proficiency and performance, and b) to explore relationships between job per-

formance ratings and selected rater/ratee factors that may influence evaluations of performance. Peer and supervisor raters were considered separately, since rating source could affect obtained relationships.

## Method

### Subjects

Participants in the research included 805 first term soldiers in five Army jobs; 172 infantrymen (MOS 11B), 168 armor crewman (MOS 19E), 148 radio teletype operators (MOS 31C), 156 light wheel vehicle mechanics (MOS 63B), and 161 medical care specialists (MOS 91A). For each job Table 1 presents the total number and average number of peer and supervisor raters per soldier ratee. In the first term soldier sample, 88.5% were male and 11.5% were female; 28% were black, 3% were hispanic, 64% white, and 5% other. Of the supervisor raters, 35% were black, 10% were hispanic, 51% were white, and 4% other.

Table 1

Summary of Supervisor and Peer Rater Assignments by Army Job

| Raters | Infantry | Armor Crewman | R-T Operator | Mechanic | Medic |
|---|---|---|---|---|---|
| Number of supervisor raters | 149 | 161 | 134 | 144 | 156 |
| Avg. no. of supervisor raters/ratee | 1.83 | 1.66 | 1.66 | 1.76 | 1.60 |
| Number of peer raters | 172 | 163 | 123 | 129 | 158 |
| Avg. no. of peer raters/ratee | 3.02 | 2.98 | 2.50 | 2.13 | 3.08 |

Instruments

The first step in this research was to develop rating scales to measure (a) performance on all relevant job factors and overall job performance in each of the five jobs, and (b) ratee personal characteristics and components of the rater-ratee relationship. In addition, a job knowledge test and a hands-on job sample test for 15 critical tasks were developed for each of the five jobs.

Job performance rating scales. Critical incident workshops were conducted with Non-Commissioned Officers (NCO), first-line supervisors for each of the target jobs. The numbers of NCO's contributing critical incidents and the numbers of incidents generated were, respectively: 11B, 51 NCOs, 906 incidents; 19E, 43 NCOs, 798 incidents; 31C, 45 NCOs, 830 incidents; 63B, 49 NCOs, 882 incidents; and 91A, 42 NCOs, 783 incidents. Rating scales were developed for each job using a variant of the behaviorally anchored rating scale procedure (Smith & Kendall, 1963). These procedures resulted in seven to ten 7-point behavior summary scales (Borman, 1979) for each of the five jobs. In addition, a 7-point summary rating of overall job performance was included on the rating form. Scores on the overall job performance rating scale were averaged across peer raters and separately across supervisor raters. This aggregate performance measure provided the primary dependent variable for this research.

Ratee traits and interpersonal relationship rating scales. Past research and conceptual considerations led to the selection of items for these scales. In addition, 25 NCOs and Comissioned Officers were interviewed to elicit ideas about factors potentially affecting job performance ratings. On the basis of these interviews and the research literature, rating scales were developed to measure several ratee traits and interpersonal relationship factors that might influence ratings. These factors included: (a) friendship

between rater and ratee (for peers), (b) trust and support between rater and ratee, (c) ratee interpersonal skills (e.g., social skills), and (d) other characteristics of ratees (e.g. likeability, moodiness) which may influence evaluation by affecting the performance-related image raters have of ratees. Responses to each rating scale in this set of measures were averaged across peer raters and separately across supervisor raters to provide an aggregate assessment of each ratee.

The composite measure of supervisor perceptions of trust and support from the ratee was composed of four items adapted from the work of Graen and his associates (e.g., Dansereau, Graen, & Haga, 1975). The four items were: (a) I can trust and depend on this soldier, (b) this soldier is willing to support me and back me up if I need it, (c) this soldier gives me help and support in getting the job done, and (d) this soldier supports and defends my decisions even if I am not there to do it myself. Responses to each item were made on a 5-point scale from strongly disagree (1) to strongly agree (5).

Responses to four items were summed to provide a measure of peer perceptions of ratee trust and job-related and emotional support. The four items were: (a) I can count on this soldier to back me up if I really need it, (b) I can trust and depend on this soldier, (c) if I had a personal problem I would feel free to discuss it with this soldier, and (d) if I had trouble performing a task, I could go to this soldier for good technical advice. Responses to each item were made on a 5-point scale from strongly disagree (1) to strongly agree (5).

The measure of rater-ratee friendship (for peers) was based on a measure used by Love (1981). Raters were asked to report friendship with the ratee on a 4-point scale with close friends (4), friends, but not close friends (3), we get along ok (2), and we don't get along with each other (1).

Other ratee attributes assessed were (a) good sense of humor, (b) likeable, (c) know-it-all, (d) "bootlicker" (for peers), (e) irritable, (f) moody, (g) rude, (h) complaining, (i) friendly, (j) socially skilled, (k) mean (for peers), and (l) boastful (for peers). Raters indicated how accurately each adjective described the ratee by using a 5-point scale from extremely inaccurate (1) to extremely accurate (5). Definitions of each attribute were provided to assist raters in making their judgments.

Hands-on, task proficiency tests. For each of the jobs, 15 critical tasks representative of the entire task domain were the target for test development work. Task proficiency tests were prepared for each of the tasks. Each task had several performance steps and each step was scored pass or fail. A proportion-passed score was derived for a soldier on each task and the proportions were averaged across tasks to yield an overall hands-on test score.

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

Cognitive ability test. Prior to entrance into military service, ratees were administered the Armed Services Vocational Aptitute Battery (ASVAB). The ASVAB is composed of 10 subtests and is used for selection and occupational classification. A composite measure of four ASVAB subtests known as the Armed Forces Qualification Test (AFQT) was used as a measure of general cognitive ability.

## Procedure

Peer and supervisor raters were trained to use the behavior-based rating scales. With reference to the rater training literature (e.g., Bernardin & Pence, 1981; Pulakos, 1984), the training can be characterized as a combination of psychometric error and frame-of-reference program. The administrator described halo, restriction-of-range, and other rating errors in lay terms and provided guidance on how to reduce those errors. Also, the logic of the behavior-based scales was carefully explained, and raters were urged to study and then properly use the behavioral anchors to arrive at their ratings. After training, peer and supervisor raters in separate groups of 3-15 made their evaluations on the job performance scales. In addition, peer and supervisor raters evaluated ratees on the interpersonal characteristics rating scales. The first term soldier (ratees) were also administered the job knowledge and hands-on job sample tests.

## Results

Table 2 presents correlations between the various ratee social traits and relationship factors and overall performance ratings by peers and supervisors. Of the interpersonal relationship factors, perceived trust and support from the ratee correlates highest with job performance ratings. Rated more highly is the performance of those soldiers who are perceived as willing to back up the rater, being someone he/she can trust and depend on, and for supervisor raters as someone who supports his/her decisions. Perceptions of being moody, irritable, a "bootlicker" (for peers) and a know-it-all are not correlated highly with job performance ratings. One finding of interest here is that the pattern of correlations with overall job performance ratings is

Table 2

Correlations Between Interpersonal/Relationship Factors and Overall Job Performance Across the Five Jobs

| Measure | Supervisory Ratings of Job Performance (N ≈ 650 Rates) | Peer Ratings of Job Performance (N ≈ 700 Rates) |
|---|---|---|
| Sense of Humor | 31 | 27 |
| Generous | -- | 25 |
| Irritable* | 22 | 15 |
| Mean* | -- | 16 |
| Bootlicker* | -- | 05 |
| Know-it-all* | 14 | 06 |
| Friendly | 22 | 27 |
| Moody* | 17 | 08 |
| Complaining* | 24 | 26 |
| Boastful* | -- | 17 |
| Likeable | 33 | 35 |
| Rude* | 20 | 26 |
| Support/Mutual Trust | 47 | 51 |
| Friendship | -- | 29 |

Note. *These variables are reverse scored (e.g., low scores on "Irritable" means the ratee is rated as very irritable).

highly similar for supervisors and peers. Within the set of interpersonal/relationship measures used by both rating sources, the rank order correlation of their respective correlations with the job performance rating was .91. We should point out, as well, that this result was not due to different amounts of measurement error associated with the interpersonal/relationship scales. The rank order correlation between the interrater reliability of each interpersonal/relationship factor and their respective correlation with the job performance rating was .12 for supervisors, and -.08 for peers.

Correlations between hands-on job knowledge test scores, cognitive ability, months in present unit, selected interpersonal/relationship factors and overall job performance ratings are presented in Tables 3 and 4. Results are shown separately for peer and supervisor raters across the five jobs. As can be seen, supervisor and peer job performance ratings show positive but low correlations with task proficiency test scores, job knowledge, and cognitive ability.

The extent of between-job variation in correlates of overall job performance ratings was also examined. Table 5 depicts correlations by job and rating source between selected variables and the overall job performance rating. A mean correlation across five jobs was computed by weighting each correlation by its associated sample size. Meta-analysis techniques (Hunter, Schmidt, & Jackson, 1982) were applied to each set of correlations to determine the extent of non-artifactual variance around the average correlation. As can be seen in Table 5, correlates of overall job performance ratings do vary somewhat across jobs, but much of the variation is attributable to sam-

Table 3

Intercorrelations Among Selected Variables Across the Five Jobs: Peer Raters

| Measures | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1. Cognitive ability | --- | | | | | | |
| 2. Task proficiency | .24 | --- | | | | | |
| 3. Job knowledge | .44 | .41 | --- | | | | |
| 4. Months in unit | -.01 | .01 | -.12 | --- | | | |
| 5. Trust and support | .05 | .10 | .09 | .01 | --- | | |
| 6. Likeable | .02 | .06 | .05 | .03 | .62 | --- | |
| 7. Overall job performance | .08 | .11 | .18 | .06 | .51 | .35 | --- |

Note. ($N \simeq 700$ ratees).

Table 4

Intercorrelations Among Selected Variables Across the Five Jobs: Supervisor Raters

| Measure | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1. Cognitive ability | --- | | | | | | |
| 2. Task proficiency | .24 | --- | | | | | |
| 3. Job knowledge | .44 | .41 | --- | | | | |
| 4. Months in unit | -.01 | .01 | -.12 | --- | | | |
| 5. Trust and support | .06 | .08 | .14 | .00 | --- | | |
| 6. Likeable | -.01 | .01 | -.02 | .07 | .10 | --- | |
| 7. Overall job performance | .07 | .15 | .23 | .14 | .47 | .33 | --- |

Note. ($N \simeq 650$ ratees).

Table 5

Correlations of Selected Variables with Ratings of Job Performance by Army Job

| Measure | Infantry | Armor Crewman | E-T Operator | Mechanic | Medic | $\bar{r}$ | % Variance Unexplained[a] |
|---|---|---|---|---|---|---|---|
| **Correlations with supervisor ratings** | | | | | | | |
| 1. Cognitive Ability | .02 | .02 | .00 | .02 | .17 | .05 | ---- |
| 2. Hands-on Tests | .33 | .04 | .11 | .30 | .16 | .19 | 32 |
| 3. Job Knowledge | .30 | .27 | .21 | .18 | .19 | .23 | ---- |
| 4. Months in Company | .01 | .13 | .16 | .20 | .29 | .16 | 27 |
| 5. Race of Ratee[b] | .13 | .13 | .04 | -.04 | -.02 | .05 | ---- |
| 6. Trust and Support | .60 | .46 | .51 | .47 | .39 | .48 | 6 |
| 7. Likeable | .40 | .29 | .41 | .28 | .35 | .34 | ---- |
| **Correlations with peer ratings** | | | | | | | |
| 1. Cognitive Ability | .07 | .01 | .16 | .08 | .07 | .07 | ---- |
| 2. Hands-on Tests | .30 | .16 | .02 | .11 | .12 | .15 | 18 |
| 3. Job Knowledge | .15 | .17 | .27 | .24 | .08 | .17 | ---- |
| 4. Months in Company | .10 | .06 | .17 | .16 | .08 | .06 | 38 |
| 5. Race of Ratee[b] | .06 | .09 | .08 | .06 | -.10 | .04 | ---- |
| 6. Trust and Support | .52 | .54 | .47 | .55 | .54 | .53 | ---- |
| 7. Friendship | .41 | .30 | .22 | .28 | .31 | .31 | ---- |
| 8. Supervisor Rating of Job Performance | .52 | .42 | .63 | .62 | .55 | .55 | 39 |

[a] Percent variance unexplained after removal of expected variance due to sampling error.

[b] Dichotomous variable (1=White/0=Black).

pling error. Correlations between supervisory and peer ratings of overall job performance were quite high across the five jobs ($r$=.43-.63). In addition, the race of ratee showed low correlations with job performance ratings.

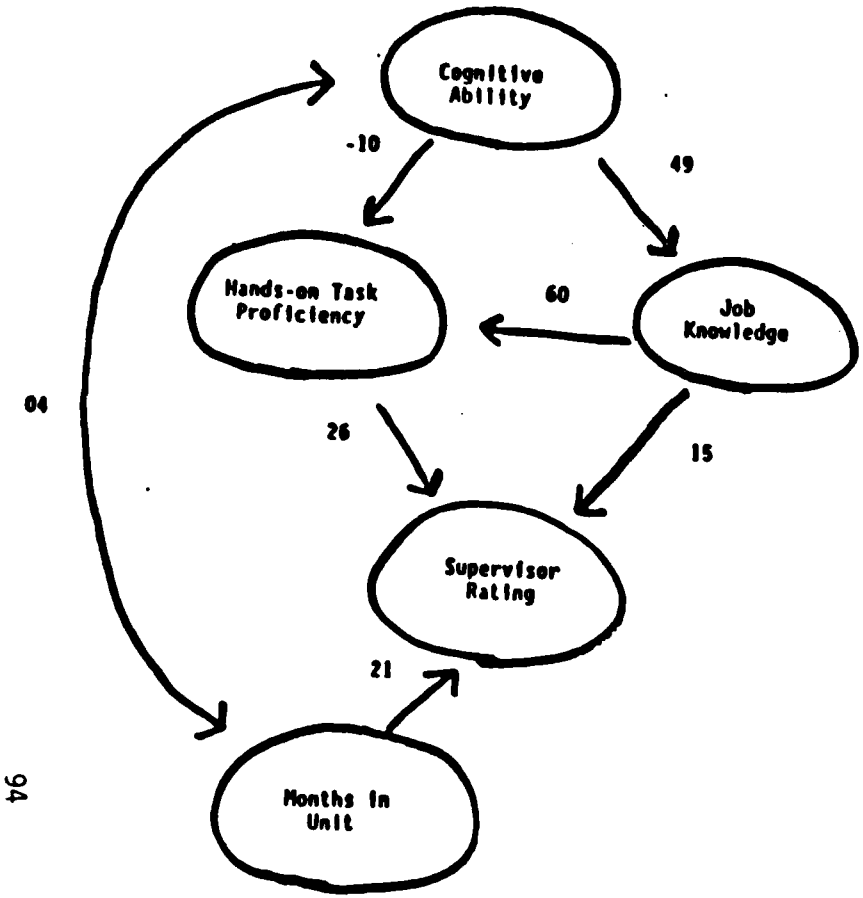
Path analytic model of ratings

Path analysis was used to examine relationships between job performance ratings and some of the measured variables potentially relevant to ratings. Hunter (1983) investigated causal relations among measures of general mental ability, job knowledge, and job performance as measured by hands-on test scores and supervisor ratings. The model examined here differed from the Hunter work by including months in the unit as an exogenous variable linked to ratings.

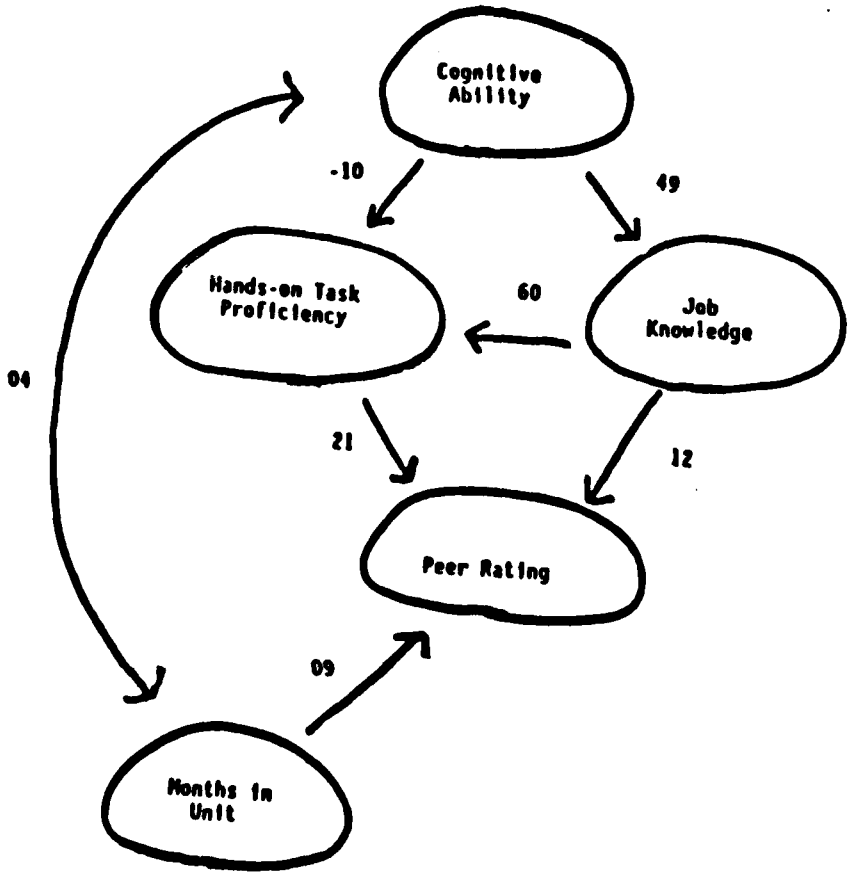
One concern here is that failure to correct observed correlation coefficients for measurement error may cause path coefficients to be biased in unknown directions (James, Mulaik, & Brett, 1982). To eliminate possible effects of measurement error on estimates of path coefficients, correlations between each pair of variables were corrected for attenuation.

Reliability information was available for all but one of the variables in the path model. The reliability estimates used to make the corrections are as follows: Cognitive ability = .90, job knowledge = .90, hands-on task proficiency = .50, supervisor ratings = .65, and peer ratings = .60. The first three estimates are internal consistency reliabilities. Interrater agreement was used to estimate the reliability of job performance ratings. The reliability of ratees' report of "months in present unit" was assumed to be 1.00.

Figure 1 presents the model of peer and supervisory ratings tested for




Supervisor Rating




Peer Rating

Figure 1. Causal model that fits the peer and supervisor rating data.

for relationships in thi research. Differences between the original correlations and correlations reproduced using the model were all less than .07, indicating a reasonably good fit for the path model. Values of the chi-square goodness of fit test were relatively low, with $\chi^2(3, N=650=10.66$, $p=.014$, for the model of supervisory ratings, and $\chi^2(3, N=700)=5.43$, $p=.143$, for peer assessments.

The models of peer and supervisory ratings presented in Figure 1 generally cooborate and extend the work of Hunter (1983). In the models, ratings are positively related to job knowledge, task proficiency, and job experience in the unit. Task proficiency had a stronger effect than job knowledge on both supervisory and peer ratings of job performance. General ability influenced job performance indirectly by contributing to the acquisition of job knowledge. For the models presented here, the multiple correlation for the prediction of overall job performance ratings was .42 for supervisory appraisals and .30 for peer assessments.

## Discussion

Little previous research has examined possible effects of ratee social traits and rater-ratee relationship factors on performance ratings (e.g., Landy & Farr, 1980; Guion, 1983). The present research addressed these shortcomings by including as measures several interpersonal and rater-ratee relationships factors, and utilizing a policy capturing framework to begin to understand the possible effects of these factors on performance judgements.

Of the factors examined, perceived trust and support from the ratee was

consistently the most important, across both different jobs and supervisor and peer rating sources. Almost 25 years ago, Kipnis (1960) emphasized the importance of supportive behavior by subordinates as a reliable basis for supervisory ratings. Kipnis proposed that supportive behaviors are the very ones which facilitate the work flow. More recent research by Graen and his associates (e.g., Graen & Cashman, 1975; Graen, Novack & Sommerkamp, 1983; etc.) indicates that the development of high levels of mutual trust and support between a supervisor and his/her subordinate creates conditions for effective performance. Subordinates who gain the trust of their superiors are likely to benefit from this close association by having more opportunities to practice skills and by receiving more individual encouragement and attention.

Peer ratings of job performance were likewise highly correlated with perceptions of emotional and job-related support from the ratee. It seems likely that soldiers who are supportive of their peers may have this support reciprocated. Further, positive correlations ($\underline{r}$ = .36 across the five jobs) between supervisor and peer ratings of trust suggest that a soldier who earns the trust of his/her peers is also supportive of his superiors. This network of relationships is likely to benefit the soldier in performing his/her job and in coping with Army life (e.g., Parker & DeCotiis, 1983).

An intriguing finding was that supervisors and peers seem to be influenced by similiar patterns of interpersonal and relationship factors. This result may be somewhat idiosyncratic to the present situation where supervisors often work closely with their troops and might tend to have perspectives on performance similar to peers. The generality of findings reported

here should be investigated in other samples.

It would have been desirable to include the rater-ratee relationship constructs and ratee traits in the same path analysis with the job proficiency and job experience measures. However, shared method variance between interpersonal relationship ratings and job performance ratings would have made comparisons of these path coefficients with across method coefficients impossible to interpret (Billings & Wroten, 1978). Thus, the path analysis work was conducted separately from the correlational, policy-capturing analysis. Comparing correlations between ratee traits and relationship factors and job performance ratings does seem legitimate in that method variance/halo is essentially equated in these comparisons.

The path analytic work presented here suggests that ratings are measuring aspects of effectiveness largely different from those assessed by task proficiency or job knowledge tests. Hands-on and knowledge tests are of course maximum performance measures and tap the "can do" part of job performance. Ratings should be measuring more the "will do", typical performance-over-time facets of performance, the motivation-related, larger term aspects of effectiveness on a job.

This is confirmed in part by strong path coefficents between cognitive ability and job knowledge, but much weaker links between ability and job performance ratings. If the opposite finding emerges when non-cognitive predictors of performance are included (e.g., temperament, vocational interest, background variables); that is, non-cognitive predictors relate more strongly to ratings and less strongly to hands-on/job knowledge criteria, then we will have more definitive evidence supporting this line of

reasoning. Data now being gathered will allow these comparisons. In sum, we believe that a better understanding of what criterion indexes are measuring may come from examining relationships between criterion measures and between different kinds of predictors and each criterion element.

In our judgment, the analysis of criterion measures does not come down to a question of which criterion is better. Rather, they each appear to be measuring somewhat different elements of job performance, each element probably important in its own right. Regarding this argument, the notion of applying multitrait-multimethod strategies to multiple criteria is compelling, and it has been accomplished to a limited extent (e.g., Lawler, 1967). However, this approach needs to take account of the possibility that different methods are measuring at least partially different criterion elements, each with some degree of validity. Accordingly, we obtain better coverage of criterion performance by employing multiple measures, provided of course that each of the measures is tapping important facets of performance.

To sum up, the present research has shed more light on some of the factors influencing supervisor and peer ratings. Correlational analysis revealed that perceived support and trust between rater and ratee is an important component of supervisory and peer ratings. Further, supervisor and peers seemed to focus on very similar factors when making performance judgments. This may help to explain the reasonably high correlations (mostly in 50's) between supervisor and peer job performance ratings in the present research.

The pattern of path coefficients also showed that task proficiency and job knowledge are significant factors in performance ratings, but for the

most part different methods of measuring job performance yield quite different results. Examining relationships between various kinds of predictor measures and the criteria should provide additional clues as to what these criteria are measuring.

References

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

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

Billings R. S., & Wroten, S. P. (1978). Use of path analysis in Industrial/Organizational Psychology: Criticisms and suggestions. Journal of Applied Psychology, 63, 677-688.

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

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

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

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

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating results across studies. Beverly Hills: Sage Publications.

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

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal analysis: Assumptions, models, and data. Beverly Hills: Sage Publications.

Kipnis, D. (1960). Some determinants of supervisory esteem. Personnel Psychology, 13, 377-391.

Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87, 78-107.

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

Parker, D. F., & DeCotiis, T. A. (1983). Organizational determinants of job stress. Organizational Behavior and Human Performance, 32, 160-177.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

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

Zedeck, S., & Kafrey, D. Capturing policies for processing evaluation data. Organizational Behavior and Human Performance, 18, 269-294.

# CRITERION REDUCTION AND COMBINATION VIA A PARTICIPATIVE DECISION-MAKING PANEL

John P. Campbell James H. Harris
Human Resources Research Organization

August 1985

Presented at a symposium, "Building a Composite Measure of Soldier Performance," at the Annual Meeting of the American Psychological Association, Los Angeles, California

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

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

# Criterion Reduction and Combination via a Participative Decision Making Panel

John P. Campbell

James H. Harris

Human Resources Research Organization

The general problem of criterion combination has been a perennial issue in personnel selection research. It is frequently the case that more than one criterion measure is available but it is the validity of one decision (e.g., select/not select) that must be established. A number of solutions to the problem have been suggested. Weighting component scores by their reliabilities, weighting in proportion to their factor loading on the general factor, weighting by each component's judged importance for the organization's goals, or equal weighting via standard scores have all been proposed. However, if the content of the criterion is taken as a definition of what the organization wants its people to be able to contribute, then the most salient procedure is to weight components in terms of their judged importance. Such judgments must be tempered by whatever psychometric data are available concerning the reliability and construct validity of each criterion component. Consequently, in the end, a number of parameters must be taken into account before a final decision is made about how to use a particular component in an overall criterion composite, and there is no analytic solution to which we can appeal. It is a complex judgment process.

This paper is about what happens when criterion data are analyzed, evaluated, and interpreted by a committee; or in this case, when a tense group of concerned psychologists attempts to evaluate and interpret a mountain of field test data on a multitude of job performance measures and reach a consensus on the nature of the criterion space and how it should be measured. It was at once an exercise in psychometrics, team building, organization development, and crises management.

The data were collected as part of a project entitled, "Improving the selection, classification, and utilization of Army enlisted personnel" (or Project A for short), which is funded through the Army Research Institute for the Behavioral and Social Sciences (ARI) and, together with the ARI research staff, is being carried out by a consortium of three firms, the Human Research Organization (HumRRO), the American Institutes for Research (AIR), and Personnel Decisions Research Institute (PDRI). Project A is a 9 year project and its overall purpose is to provide the data necessary for designing improvements in the selection/classification system for enlisted personnel. The intended improvements are in the form of developing new selection and classification tests to supplement the Army's Vocational Aptitude Battery (ASVAB) and to validate all selection/classification measures against a broad array of job performance criteria. It is probably the largest R&D project ever undertaken in personnel management. A complete description can be found in Eaton and Goer (1983).

With the above as background, the specific objectives of the current paper are the following.

1) To describe the nuts and bolts of what are called the criterion field tests in Project A. The field tests were the final development step before the full scale concurrent validation of a comprehensive test battery involving over

10,000 incumbents in 19 different jobs (MOS). The field tests gathered data from 1369 people on a comprehensive array of criterion measures and the general analytic strategy and overall findings will be summarized.

2) To discuss interpretations made by the committee of concerned psychologists.

3) To outline the current working model of job performance for the domain of skilled jobs.

## Field Test: Introduction

The data upon which this paper (and the others in the symposium) are based were collected as part of the field tests for the job performance criterion measures being used in the Army's selection and classification project (Project A). The nature of this very large project has already been described. We are concerned here with the development of a set of criterion measures that cover the entire domain of job performance for the complete population of entry level skilled jobs in the Army. The objectives of the criterion field tests were to:

1) Provide item/scale analyses for the subsequent revision of the criterion measures to be used in the major Project A validation samples.

2) Provide data on the reliabilities and factor structures of the performance ratings, job sample measures, and job knowledge tests.

3) Provide data to estimate the interrelationships among the major kinds of criterion measures.

4) Evaluate the data collection procedures to be used in the large scale concurrent validation.

## The Sample

The sample for the field tests was drawn from 9 different jobs, or Military Occupational Specialties (MOS), and from six different locations. The 9 jobs and their MOS designation were as follows.

11B Infantryman

13B Cannon Crewman

19E Tank Crewman

31C Radio Operator

63B Vehicle and Generator Mechanic

64C Motor Transport Operator

71L Administrative Specialist

91A Medical Care Specialist

95B Military Police

Tables 1 and 2 provide a breakdown of the sample sizes by MOS and by location. USAREUR refers to the data collection site just outside Frankfurt, Germany.

Table 1

Soldiers by MOS by Location

| LOCATION | MOS | | | | | | | | | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|
| | 11B | 13B | 19E | 31C | 63B | 64C | 71L | 91A | 95B | |
| Fort Hood | | | | | | | 48 | | 42 | 90 |
| Fort Lewis | 29 | | 30 | 16 | 13 | | | 24 | | 112 |
| Fort Polk | 30 | | 31 | 26 | 26 | | 60 | 30 | 42 | 245 |
| Fort Riley | 30 | | 24 | 26 | 29 | | 21 | 34 | 30 | 194 |
| Fort Stewart | 31 | | 30 | 23 | 27 | | | 21 | | 132 |
| USAREUR | 58 | 150 | 57 | 57 | 61 | 155 | | 58 | | 596 |
| TOTAL | 178 | 150 | 172 | 148 | 156 | 155 | 129 | 167 | 114 | 1369 |

Table 2

Soldiers by Sex by Race

| Race | Sex | | |
|---|---|---|---|
| | Male | Female | Total |
| Black | 330 | 58 | 388 |
| Hispanic | 37 | 3 | 40 |
| White | 789 | 104 | 893 |
| Other | 43 | 5 | 48 |
| Total | 1199 | 170 | 1369 |

## The Criterion Measures

The general model and procedures for criterion development in Project A have already been described in the other papers in this symposium. The basic cycle of a comprehensive literature review, conceptual development, scale consiruction, pilot testing, scale revision, field testing and proponent (management) review was followed for each kind of criterion measure. The primary goals of criterion measurement in Project A were to: a) make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency, b) compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e. a multi-trait, multi-method approach), c) develop rating scale measures of performance factors that are common to all first tour enlisted MOS (army-wide measures), d) develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance, and e) evaluate existing archival and administrative records as possible indicators of job performance.

Given these intentions, the overall criterion development effort focused on three major methods: hands-on job samples, multiple choice knowledge tests, and ratings. The behaviorally anchored rating scale (BARS) procedure was extensively used in the development of the rating methods. A brief description of each of the major types of measures is given below.

Hands-On Measures. A comprehensive task sampling procedure was used to define the population of tasks in each MOS and then select 30 job tasks to represent it. After the 30 tasks per MOS were selected, the two major development tasks that remained before actual preparation of tests were the review of the task lists by the proponent schools and the assignment of

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

The completeness and representativeness of the task lists were officially reviewed by the proponent school. Only slight changes were made in the task lists as a result of the reviews.

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

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

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

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

measures and scorer directions were then pretested on 5 to 10 incumbents in each MOS and revised accordingly.

MOS-Specific Job Knowledge Tests. Concurrently, a paper-and-pencil, multiple-choice job knowledge test was developed to cover all of the 30 tasks in the MOS lists. The item content was generated on the basis of training materials, job analysis information, and interviews, with 3 to 16 items prepared for each of the 30 tasks. For the 15 tasks also measured hands-on, the knowledge items were intended to be as parallel as possible to the steps that comprised the hands-on mode. The knowledge tests were pilot tested on approximately 5 job incumbents per MOS. After revision they were deemed ready for tryout with the field tests samples.

MOS Specific Task Ratings. A seven point rating scale was also developed for each of the 15 job tasks that were measured hands-on.

MOS-Specific BARS Scales. Following the procedure for developing behaviorally anchored rating scales, four critical incident workshops involving 70-75 officers and NCO's were completed for each of the 9 MOS. The usual retranslation step was carried out, and six to nine MOS-specific performance rating scales (Behaviorally Anchored Rating Scales, BARS) were developed for each MOS. Directions and training materials for scales were also developed and pretested.

Army-wide Rating Scales. Army-wide measures are defined as measures of performance factors that are defined in the same way across all MOS. To construct rating scales of army-wide factors, six critical incident workshops involving 77 officers and NCO's were conducted. On the basis of the critical incidents collected in all workshops, a preliminary set of 15 Army-wide performance dimensions was identified and defined. Using a combination of workshop and mail survey participants (N = 61), the initial set of dimensions was retranslated and 11 Army-wide performance factors survived. The scaled cri-

tical incidents were used to define anchors for each scale, and directions and training materials for raters were developed and pretested.

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

Administrative (Archival) Indices. A major criterion development effort was a systematic comparison of information found in the computerized Enlisted Master File (EMF), the Official Military Personnel File (OMPF), which is a series of several microfiche files kept centrally at the military personnel center, and the Military Personnel Records Jacket (201 File), which is the primary hand copy file that stays with the individual. A sample of 750 incumbents, stratified by MOS and by location, was selected and the files searched. For the 201 Files the research team made on-site visits and used a previously developed protocol to record the relevant information. A total of 14 items of information, including awards, letters of commendation, and disciplinary actions, seemed, on the basis of their base rates and judged relevance, to have at least some potential for service as criterion measures.

Unfortunately, the microfiche records appeared too incomplete to be useful and searching the 201 Files was cumbersome and expensive. Consequently, it was decided to try out a self-report measure for the 14 administrative indices and compare it to actual 201 File information for the people in the field trials.

Training Achievement Tests. For each MOS an item budget was prepared matching job duty areas to course content modules and specifying the number of items that should be written for each combination. An item pool that reflected the item budget was then written by a team of SME's contracted for that purpose.

Next, training content SME's and job content SME's judged each item in

terms of its relevance for training, its importance for the job (under each of three scenarios - current peacetime, full alert, and a European conflict - in a repeated measures design), and its difficulty. The items were then "re-translated" back into their respective duty areas by the job SME's and into their respective training modules by the training SME's. Items were designated as "job only" if they reflected task elements that were described as an important part of the job but had no match with training content; such items are intended to be a measure of incidental learning in training.

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

Field Test Criterion Battery

The complete array of specific criterion measures that was actually used at each field test site is given below. For each rating scale every effort was made to obtain a complete set of supervisor, peer, and self ratings. Also, the following distinctions are relevant for interpreting this list of variables. First, the distinction between MOS-specific and Army-wide is that the Army-wide measures are the same across all MOS. That is, the same questionnaire or the same rating scale is used. The content of the MOS specific measures, regardless of whether they are job samples, knowledge tests, or ratings is specific to a particular job and is based on the task content of that job. Second, the total sample of tasks is divided into common tasks and unique tasks. Common tasks (e.g., first aid, use of chemical/biological warfare gear, etc.) are taught in basic training and are the responsibility of every enlisted man or woman. Unique tasks are only taught in context of a specific MOS. Both kinds were sampled for each MOS with the relative pro-

portion determined by their relative judged importance in the MOS. For example, tasks designated as common tasks are relatively more crucial for the combat specialties. Finally the judgment (i.e. rating) of "NCO potential" refers to the rating of a first tour enlisted man or woman's potential for being an effective non commissioned officer, with supervisory responsibilities, during the second tour of duty, assuming the individual would reenlist.

A. MOS-Specific Performance Measures

1) Paper-and-pencil tests of knowledge of task procedures consisting of 3-16 items for each of 30 major job tasks for each MOS. Item scores can be aggregated in at least the following ways:

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

2) Hands-on measures of proficiency on tasks for each MOS when the 15 tasks were selected from the 30 tasks measured with the paper-and-pencil test.

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

3) Ratings of performance, using a 7 point scale, on each of the 15 tasks measured via hands-on methods by:

   - Supervisors
   - Peers
   - Self

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

   - Supervisors
   - Peers
   - Self

5) A general rating of overall MOS task performance by:

   - Supervisors
   - Peers
   - Self

6) A job history questionnaire administered to incumbents to determine the frequency and recency of task performance on the 30 tasks being measured.

B. Army-Wide Measures

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

- Technical Knowledge/Skill
- Initiative/Effort
- Following Regulations/Orders
- Integrity
- Leading and Supporting
- Maintaining Assigned Equipment
- Maintaining Living/Work Areas
- Military Appearance
- Physical Fitness
- Self-Development
- Self-Control

2) A rating of general overall effectiveness as a soldier by:

- Supervisors
- Peers
- Self

3) A rating of non-commissioned officer (NCO) potential by:

- Supervisors
- Peers
- Self

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

- Supervisors
- Peers
- Self

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

6) The same administrative indices taken from 201 Files.

7) The Environmental Qeustionnaire which is a 44 item descriptive questionnaire completed by both incumbents and supervisors for the purpose of describing 14 factors pertaining to organizational climate, structure, and practices.

8) A Leader Behavior Questionnaire designed to permit incumbents and and supervisors to describe leadership policies and practices in the unit,

9) A "measurement method "questionnaire designed to elicitopinions about the fairness and face validity of each criterion measure.

## Procedure

For the purpose of data collection in the field tests the criterion measures were divided into four major blocks corresponding to:

1) Hands-on (job sample) measures - (HO).

2) Rating measures (R) - both army-wide and MOS specific.

3) Paper-and-pencil tests of job knowledge ($KN_5$).

4) Paper-and-pencil measures of training achievement ($KN_3$).

Each block comprised one half day of participant time and each participant was tested for a two day period. During the week preceding data collection at each research site the scorers for the hands-on (job sample) measure were given two days of training on scoring procedures, their influence on the reliability and validity of the measures, and the overall design and nature of Project A.

The major steps in the procedure were as follows.

### Advance Preparation on Site

This activity required approximately three days per test site for:

1) briefings to Commanders of the units supplying the troops to clarify the test objectives, activities, and requirements,

2) examination of the test site, equipment, supplies and special requirements for the data collection and set-up of the hands-on test stations,

3) training of the test administrators and scorers, and

4) a dry-run of the test procedures.

An officer and two NCO's from one of the supporting units was assigned to support the field test. The officer provided liaison between the data collection team and the tested units; and the NCO's coordinated the flow of equipment and personnel through the data collection procedures. The logistics plan and test schedule were reviewed with the unit's administrative staff, after which civilian and military scorers and other data personnel were trained. In the training phase, a dry run of the procedures followed the data collection schedule and used the personnel and locations designated for the test. The training focused on the handling of problem situations, particularly those requiring remediation by the scientific staff.

Each test site had a test site manager (TSM) who supervised all of the research activity and maintained the orderly flow of personnel through the data collection points.

Scorer Training. Training for scorers and a dry run of the test procedures for the hands-on (HO) measures proceeded as follows:

Two project staff members conducted the training for each MOS. Training began with an orientation session for the scorers . Then, staff members reviewed five HO tasks with the scorers by describing the equipment/material requirements, the procedures for setting up testing stations, and the specific instructions for administering and scoring each HO test. The scorers then alternated evaluating each other performing the tasks. This provided experience in administering the HO tests and scoring the performance measures of each. Project staff coached the "performers" to make unusual, as well as common, incorrect actions in order to give scorers practice in detecting and recording errors. The above procedure also identified the steps of tasks that differ because of local Standard Operating Procedures (SOP). When so identified, allowances were made for such differences in the test instructions.

The second day of training was devoted to a dry-run of the test procedures. All scorers simultaneously evaluated a staff member performing a task. Problems arising from the instructions, test procedures, or task steps were identified and corrected.

## Administration of the Measures

The administration schedule for a typical site (Fort Stewart, Georgia) is shown in Figure 1. The field test proceeded as follows: Thirty 31C and thirty 19E soldiers arrived at the test site Thursday 21 Feb 85 at 0745. Each MOS was divided randomly into two groups of 15 soldiers each, identified as Groups A, B, C, or D. Each group was directed to the appropriate area to begin the administration appropriate for that group. They rotated under the direction of the test site manager (TSM) through the scheduled areas according to the schedule shown in Figure 1. The sequence was repeated for 30 91B and 30 63B soldiers beginning Monday, 25 Feb 85 and for 30 11B soldiers on Wednesday 27 Feb 85. The order of administration of the measures was counterbalanced.

Figure 1

Field Test Schedule for Fort Stewart
19 Feb - 28 Feb 85

| Group[1] | 31C A B | 19E G H | 91A E F | 63B C D | 11B I J |
|---|---|---|---|---|---|
| Tuesday 19 Feb 85 | ----------Scorer Training (All Scorers)---------- | | | | |
| Wednesday 20 Feb 85 | ----------Scorer Training (All Scorers)---------- | | | | |
| Thursday AM | PH PK$_5$ | PK$_5$ | PK$_3$ | | |
| 21 Feb 85 PM | K$_3$ H | R | R | | |
| Friday AM | RS K$_3$S | H | K$_5$ | | |
| 22 Feb 85 PM | MK$_5$ MR | MK$_3$S | MHS | | |
| Monday AM | | PH | PK$_5$ | PK$_5$ | PK$_3$ |
| 25 Feb 85 PM | | K$_3$ | H | R | R |
| Tuesday AM | | RS | K$_3$S | H | K$_5$ |
| 26 Feb 85 PM | | MK$_5$ | MR | MK$_3$S | MHS |
| Wednesday AM | | | | PH | PK$_5$ |
| 27 Feb 85 PM | | | | K$_5$ | H |
| Thursday AM | | | | K$_3$S | R |
| 28 Feb 85 PM | | | | MR | MK$_3$S |

[1] Each Group equals 15 soldiers in same MOS.

Code: P = Personal and Job History form
K$_{3;5}$ = Task 3 or Task 5 Knowledge Measures
H = Hands-on Measures
R = Self and Peer Ratings
S = Supervisor (rater and endorser) ratings
M = Measurement Method Questionaire
E = Records Extraction (201 File)

Before any instruments were administered to any soldier, each was asked to read a Privacy Act Statement, DA Form 4368-R. They were then administered the Job History and Background Information forms and given 30 minutes to complete them.

Administration of Job Samples (15 tasks measured hands-on). Depending on the task being measured, the location was outside (vehicle maintenance, weapons cleaning) or inside (measure distance on a map). Scorers assigned to each test station ensured the required equipment was on-hand, that the station was set up correctly, and followed the procedures for administering and scoring the tests. As each soldier entered the test station, the scorer read aloud the instructions to the soldier and began the measure. The length of time a soldier was at the test station depends both on the individual's speed of performance and the complexity of the task.

Training Achievement Tests. The training knowledge test for each MOS was in three booklets. The sequence of the booklets was alternated so that soldiers sitting next to each other had different booklets. The purpose of using booklets is to try to control the effects of fatigue and waning interest. Soldiers had 45 minutes for each booklet and a 10-15 minute smoke, stretch, and complain break between booklets.

Rating Scales. The administration of the peer, self, and supervisory ratings proceeded as follows:

The ratings are designed around "rating units." Each rating unit consists of the individual soldier to be evaluated, four identifiable peers, and two identifiable supervisors. A peer is defined as an individual soldier to

be evaluated who has been in the unit for at least two months and has observed the ratee's job performance on several occasions. A supervisor is defined as the individual's first or second line supervisor (normally his rater and endorser.)

The procedure for assigning ratees to raters (both peers and supervisors) consists of two major steps:

1) A screening step in which it is determined which ratees could potentially be rated by which raters; and,

2) A computerized random assignment procedure which assigns raters to ratees within the constraints that (a) the rater indicated he/she could rate the ratee (Step 1); (b) ratees with few potential raters are given priority in the randomized assignment process; (c) the number of ratees assigned the various raters is equalized as much as possible across raters; and (d) the number of ratees any given rater is asked to rate does not exceed a preset maximum.

The potential raters were be given an alphabetized list of the ratees. They were told the purpose of the ratings within the context of the research and the criteria, e.g., minimum length of period of working together, which they should use in deciding who they could rate. They were told the maximum number of people they would be asked to rate and that assignments of ratees to raters would be accomplised randomly. They were further told that the randomization procedure would attempt to equalize as much as possible the number of ratees that any one rater will have to rate. The importance of their careful and comprehensive examination of the list of ratees was emphasized.

Next the rating scale administrator, using the training guide, discussed the content of each effectiveness category, and urged raters to avoid common rating errors.

A major thrust of the rater training program was an attempt to standardize the rating task for raters. With the lack of control to be expected, an important concern was that all raters face the same (or a very nearly similar) rating task. A serious potential confounding involves rating unit and administrator. Lower average ratings for some rating units may be a result of different sets (i.e., "rate more severely") provided by administrators handling those rating units rather than true performance deficiencies in the rating units. Standardization of the administration helps reduce this potential problem.

A second major thrust of the rater training program was to make it possible to obtain high quality ratings from the target subjects, their peers, and their supervisors with a minimum of reading necessary on the part of the raters. This was, as much as possible, an oral administration; the rating program is not dependent on raters' reading large amounts of material.

MOS-Specific Job Knowledge Tests. The MOS-specific knowledge tests were grouped into four booklets of about seven or eight tasks per booklet. Each booklet required about 45 minutes to complete. The order of the booklets and the order of the tasks in each booklet were rotated. There was a 10-15 minute smoke, stretch, and complain break between booklets. Again, the purpose of the booklets was to try to control the effects of fatigue and waning interest. The measurement method rating was administered at the conclusion of the second day's activities.

## Analysis

The general analytic steps were straightforward and consisted of the following.

1) Item analysis for each job knowledge test for each MOS.
2) Item analysis for the training achievement tests for each MOS. An analysis of item responses was done for a sample of 50 trainees as well as for the field test samples.
3) An item analysis summary table for each knowledge test for each MOS. The table for each MOS summarized item discrimination indices, item difficulties, and the frequency of items that were flagged for various kinds of potential keying errors (e.g., negative correlation with total score, high frequency of response for incorrect answer).
4) An item (where task = item) analysis for each "hands-on" (job-sample) test.
5) A frequency distribution and scale statistics for each rating scale for each MOS.
6) Inter-rater reliabilities for the individual rating scales.
7) Split half correlations (Spearman-Brown estimates) for the knowledge tests and hands-on measures, test-retest coefficients for the hands-on measures, and internal consistency indices where applicable.
8) A complete intercorrelation matrix of all the criterion variables for each MOS down to the scale score and task score level (i.e. the matrix included all the variables listed in the previous section).
9) A reduced intercorrelation matrix that included the following variables.

a. Three scores derived from the 15 job sample tasks (i.e. hands-on).

1 - Total score on all 15 tasks
2 - Total score on common tasks
3 - Total score on MOS specific tasks

b. The supervisor/peer ratings of incumbent performance on the 15 hands-on tasks (i.e. performance ratings on the tasks that were included in the job sample).

Supervisor's Ratings

4 - Average rating on all 15 tasks
5 - Average rating on common tasks
6 - Average rating on MOS specific tasks

Peers

7 - Average rating on all 15 tasks
8 - Average rating on common tasks
9 - Average rating on MOS specific tasks

c. MOS specific job knowledge tests - each of which included items sampled from 30 major job tasks.

For the 15 taks also measured hands-on

10 - Total score
11 - Total score on common tasks
12 - Total score on MOS specific tasks

For the 15 tasks not also measured hands-on

13 - Total score
14 - Total score on common tasks
15 - Total score on MOS specific tasks

d. Training achievement tests.

16 - Total score (all items)

e. Army-wide BARS scales.

Average of the eleven individual scales developed by the BARS procedure

17 - Supervisor ratings
18 - Peer ratings

Overall performance scale

19 - Supervisor ratings
20 - Peer ratings

NCO potential scale

21 - Supervisor ratings
22 - Peer ratings

General technical knowledge and skill scale

23 - Supervisor ratings
24 - Peer ratings

Average rating on the 14 common tasks included as part of the Army-wide rating package

25 - Supervisor ratings
26 - Peer ratings

10) Factor analyses were also been carried out for the following matrices.

a) The matrix described in #9 above.

b) All rating scales.

c) A 45 x 45 multi-trait, multi-method matrix consisting of job sample (hands-on) scores, knowledge test scores, and supervisory ratings (methods) for the 15 job tasks (traits) on which performance was assessed by each of the three methods.

## RESULTS

Since all analyses were done for each of nine jobs, the number of tables that can be created quickly becomes very large. What's reported here are illustrative results for two very different jobs, infantryman (11B) and administrative specialist (71L). They should give the flavor of the results and portray the major issues that must be confronted.

### Item/Scale Analyses

The overall means and variances of a selected number of criterion variables for the two MOS are shown in Tables 3 and 4. All ratings were on a 7 point scale. The job knowledge scores are reported in terms of percent correct but the training test score is in terms of total number correct. Each hands-on task test was scored in terms of number of steps done correctly and the total score is the percentage of total steps correct across all 15 tasks.

- - - - Tables 3 and 4 About Here - - - -

It was gratifying that each measure produced considerable variance. In no case were there highly leptokurtic or skewed distributions.

At this stage of their development, the paper-and-pencil tests proved relatively difficult. For example, the means for the training achievement tests tended to be between 50 and 60 percent correct, as were the means for the job knowledge tests.

The distributions for the rating scales were surprisingly free of leniency and skewness, which perhaps illustrates again the major difference between rating measures as research instruments versus operational performance appraisals. One consistent finding relative to the ratings was that

the peer ratings had a slightly higher mean and smaller standard deviation than the supervisor ratings, but the differences were not large.

Test and Scale Reliability

The basic reliability information is summarized in Tables 5 and 6.

- - - - Tables 5 and 6 About Here - - - -

For the hands-on and knowledge tests, a split-half (odd-even) coefficient (Spearman-Brown Corrected) was computed while the principal index for the rating scales was an estimate of the intra class correlation. The coefficient is an estimate of the inter rater agreement between two raters (i.e. the reliability of the average of two, three, etc. raters would be correspondingly higher).

In general, with two exceptions, the inter rater agreements for the rating measures were as high or higher than those usually found for rating measures (cf. Landy & Farr). Peer ratings using 3 or more raters would yield very high reliabilities. Also, the reliabilities for the BARS scales tended to be higher than those for non-BARS scales and, as would be expected, the reliability of the average rating across several scales was greater than for a single scale. The high split half coefficients for the knowledge tests illustrate again that carefully constructed achievement tests with a large number of items are very reliable instruments.

Before the field tests, the reliabilities of the hands-on measures were an unknown quantity. There is very little previous literature on which to base an expectation. During the pilot tests there was very high agreement between scorers but it was not possible to obtain two scorers per task during

the field tests. The coefficient reported in the tables is the split half coefficient obtained by correlating the score on 8 tasks with the score on 7 tasks and correcting to a total length of 15 tasks. The reliabilities are again reasonably high, particularly in view of the fact that each task was graded by a different scorer.

For 71L one additional piece of reliability information is available. That particular MOS was one of four jobs for which there was a test-retest estimate of reliability. That is, for four MOS the participants were asked to return one week later and they were retested on the hands-on measures. The test-retest coefficient was .69.

Table 3

Means and standard deviations for selected criterion variables: 11B

| | N | MEAN | STANDARD DEVIATION |
|---|---|---|---|
| 1) HO: Total Score | 162 | 56.1 | 12.3 |
| 2) Av Rating: HO Tasks (Supv) | 149 | 4.5 | 0.6 |
| 3) Av Rating: HO Tasks (Peer) | 171 | 4.5 | 0.6 |
| 4) Job KN Test Score (all HO) | 172 | 54.1 | 11.4 |
| 5) Job KN Test Score (non HO) | 172 | 59.5 | 10.9 |
| 6) Training KN Test Total Score | 166 | 87.4 | 20.1 |
| 7) Avg BARS (Supv) | 149 | 4.5 | 0.8 |
| 8) Avg BARS (Peer) | 172 | 4.5 | 0.7 |
| 9) BARS: Overall (Supv) | 149 | 4.5 | 1.0 |
| 10) BARS: Overall (Peer) | 172 | 4.6 | 0.8 |
| 11) BARS: NCO Potential (Supv) | 149 | 4.0 | 1.4 |
| 12) BARS: NCO Potential (Peer) | 171 | 4.1 | 1.1 |
| 13) BARS: Tech Skill (Supv) | 149 | 4.5 | 1.1 |
| 14) BARS: Tech Skill (Peer) | 172 | 4.6 | 0.9 |
| 15) Av MOS BARS (Supv) | 149 | 4.5 | 0.7 |
| 16) Av MOS BARS (Peer) | 172 | 4.5 | 0.6 |
| 17) Av Rt: 14 Comm (Supv) | 149 | 4.9 | 0.7 |
| 18) Av Rt: 14 Comm (Peer) | 171 | 5.0 | 0.6 |

Table 4

Means and standard deviations for selected

criterion variables: 71L

| | | N | MEAN | STANDARD DEVIATION |
|---|---|---|---|---|
| 1) | HO: Total Score | 126 | 62.1 | 9.9 |
| 2) | Av Rating: HO Tasks (Supv) | 99 | 5.0 | 0.7 |
| 3) | Av Rating: HO Tasks (Peer) | 55 | 5.0 | 0.6 |
| 4) | Job KN Test Score (all HO) | 128 | 59.1 | 10.8 |
| 5) | Job KN Test Score (non HO) | 127 | 51.5 | 10.2 |
| 6) | Training KN Test Total Score | 129 | 54.3 | 10.3 |
| 7) | Avg BARS (Supv) | 109 | 4.7 | 0.8 |
| 8) | Avg BARS (Peer) | 64 | 4.8 | 0.7 |
| 9) | BARS: Overall (Supv) | 109 | 4.4 | 1.2 |
| 10) | BARS: Overall (Peer) | 64 | 4.7 | 0.9 |
| 11) | BARS: NCO Potential (Supv) | 109 | 4.8 | 1.3 |
| 12) | BARS: NCO Potential (Peer) | 64 | 4.8 | 0.9 |
| 13) | BARS: Tech Skill (Supv) | 109 | 4.5 | 1.2 |
| 14) | BARS: Tech Skill (Peer) | 64 | 5.0 | 0.9 |
| 15) | Av MOS BARS (Supv) | 107 | 4.5 | 0.9 |
| 16) | Av MOS BARS (Peer) | 64 | 4.7 | 0.6 |
| 17) | Av Rt: 14 Comm (Supv) | 105 | 4.5 | 0.8 |
| 18) | Av Rt: 14 Comm (Peer) | 60 | 4.8 | 0.7 |

Table 5

Reliability estimates for selected variables: 11B

| | RELIABILITY | |
|---|---|---|
| | Split-Half | Inter rater Agreement |
| 1) HO: Total Score | 49 | -- |
| 2) Av Rating: HO Tasks (Supv) | -- | 74 |
| 3) Av Rating: HO Tasks (Peer) | -- | 77 |
| 4) Job KN Test Score (all HO) | 84 | -- |
| 5) Job KN Test Score (non HO) | 82 | -- |
| 6) Training KN Test Total Score | 91 | -- |
| 7) Avg BARS (Supv) | -- | 82 |
| 8) Avg BARS (Peer) | -- | 80 |
| 9) BARS: Overall (Supv) | -- | 64 |
| 10) BARS: Overall (Peer) | -- | 47 |
| 11) BARS: NCO Potential (Supv) | -- | 74 |
| 12) BARS: NCO Potential (Peer) | -- | 57 |
| 13) BARS: Tech Skill (Supv) | -- | 49 |
| 14) BARS: Tech Skill (Peer) | -- | 58 |
| 15) Av MOS BARS (Supv) | -- | 78 |
| 16) Av MOS BARS (Peer) | -- | 81 |
| 17) Av Rt: 14 Comm (Supv) | -- | 77 |
| 18) Av Rt: 14 Comm (Peer) | -- | 78 |

Table 6

Reliability estimates for selected variables: 71L

| | RELIABILITY | |
|---|---|---|
| | Split-Half | Inter rater Agreement |
| 1) HO: Total Score | 66 | -- |
| 2) Av Rating: HO Tasks (Supv) | -- | 75 |
| 3) Av Rating: HO Tasks (Peer) | -- | 60 |
| 4) Job KN Test Score (all HO) | 71 | -- |
| 5) Job KN Test Score (non HO) | 63 | -- |
| 6) Training KN Test Total Score | 84 | -- |
| 7) Avg BARS (Supv) | -- | 54 |
| 8) Avg BARS (Peer) | -- | 82 |
| 9) BARS: Overall (Supv) | -- | 77 |
| 10) BARS: Overall (Peer) | -- | 70 |
| 11) BARS: NCO Potential (Supv) | -- | 29 |
| 12) BARS: NCO Potential (Peer) | -- | 60 |
| 13) BARS: Tech Skill (Supv) | -- | 74 |
| 14) BARS: Tech Skill (Peer) | -- | 22 |
| 15) Av MOS BARS (Supv) | -- | 77 |
| 16) Av MOS BARS (Peer) | -- | 81 |
| 17) Av Rt: 14 Comm (Supv) | -- | 84 |
| 18) Av Rt: 14 Comm (Peer) | -- | 57 |

## Scale Intercorrelations

Project A might be accused by some of collecting a bit too much data. The accusation becomes the most credible when the intent is to calculate an intercorrelation matrix among the principal criterion measures. The list is long or short depending on how much aggregation one is willing to tolerate. If the supervisor, peer, and self ratings for all rating scales are counted, along with hands-on and knowledge test scores for each of the 30 tasks, the total number of criterion variables is 162. Thatis a few too many to interpret at a glance, without further reduction. One strategy is cluster or factor analysis but if you are lo-tech, feel anxious in the presence of too many partial or semi-partial correlations, and respect rational professional judgment, then such analyses are a court of last resort. Consequently, we first eliminated self ratings from further consideration because they lower correlations with other variables and greater leniency. Ratings for peers and supervisors were then averaged on the assumption that 2 raters approximated parallel measures. In addition, we averaged across the eleven army-wide scales, the 14 common task scales, the MOS task scales, and the MOS-specific BARS scales. For the hands-on and job knowledge tests scores were totaled for the 15 tasks measured hands-on.

After all this was done, the variable list was reduced to 10 and the intercorrelation matrices for 11B and 71L are shown as Tables 7 and 8.

- - - - Tables 7 and 8 About Here - - - -

These two matrices are illustrative of some basic truths. First, the methods correlate more highly within themselves than they do across measures. If one were to examine a multi-method (hands-on, knowledge tests, ratings) - multi-trait (the 15 tasks) matrix and submit it to a factor analysis, the factors would be defined by methods rather than job tasks. This is not unlike what happens when individual assessment center measures are factored.

Factors tend to be defined by the particular exercise or test rather than the trait (Sackett & Dreher, 1982). However, two points are crucial. The variables of task proficiency, job knowledge, and general soldiering performance are not identical but they are also not independent, in spite of the influence of method variance. Also, it is still one of the great unanswered questions in applied psychology as to whether what we refer to as method variance (e.g., halo) in ratings, paper-and-pencil knowledge tests, or job sample tests is relevant and valid, or simply noise. It is not necessarily error, but may indeed reflect individual differences in performance that are quite relevant.

Table 7. Reduced Correlation Summary: 11B

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1) Total score on training knowledge tests | -- | | | | | | | | | |
| 2) Total score on 15 hands-on tests | 40 | -- | | | | | | | | |
| 3) Total score for MOS-specific knowledge tests on 15 HO tests | 70 | 55 | -- | | | | | | | |
| 4) Avg of Sup & Peer ratings on 15 HO tasks | 33 | 41 | 35 | -- | | | | | | |
| 5) Avg of Sup & Peer ratings on MOS-specific BARS | 23 | 41 | 27 | 63 | -- | | | | | |
| 6) Avg of Sup & Peer ratings on task knowledge & skill | 27 | 36 | 30 | 55 | 65 | -- | | | | |
| 7) Avg of Sup & Peer ratings over all 11 BARS scales | 25 | 36 | 26 | 56 | 74 | 72 | -- | | | |
| 8) Avg of Sup & Peer ratings for overall performance | 19 | 32 | 23 | 51 | 68 | 63 | 74 | | | |
| 9) Avg of Sup & Peer ratings for NCO potential | 21 | 35 | 27 | 45 | 62 | 61 | 73 | 64 | -- | |
| 10) Avg of Sup & Peer ratings across common tasks | 24 | 41 | 32 | 62 | 72 | 59 | 67 | 62 | 55 | - |

Table 8. Reduced Correlation Summary: MOS 71L

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1) Total score on training knowledge tests | -- | | | | | | | | | |
| 2) Total score on 15 hands-on tests | 54 | -- | | | | | | | | |
| 3) Total score for MOS-specific knowledge tests on 15 HO tests | 63 | 52 | -- | | | | | | | |
| 4) Avg of Sup & Peer ratings on 15 HO tasks | 23 | 20 | 07 | -- | | | | | | |
| 5) Avg of Sup & Peer ratings on MOS-specific BARS | 24 | 23 | 12 | 53 | -- | | | | | |
| 6) Avg of Sup & Peer ratings on task knowledge & skill | 22 | 15 | 14 | 50 | 55 | -- | | | | |
| 7) Avg of Sup & Peer ratings over all 11 BARS scales | 19 | 17 | 12 | 45 | 60 | 63 | -- | | | |
| 8) Avg of Sup & Peer ratings for overall performance | 22 | 25 | 16 | 39 | 45 | 41 | 70 | -- | | |
| 9) Avg of Sup & Peer ratings for NCO potential | 21 | 19 | 07 | 35 | 40 | 35 | 65 | 42 | -- | |
| 10) Avg of Sup & Peer ratings across common tasks | 18 | 19 | 12 | 23 | 30 | 28 | 40 | 35 | 28 | -- |

## True Score Relationships

The inter correlations in the previous tables are between fallible scores on each variable score. To get closer to the "truth" about the criterion space, the intercorrelations were corrected for attenuation which yielded an estimate of the true score intercorrelation matrix. The matrices for 11B and 71L are shown in Tables 9 and 10.

These correlations were computed on the assumption that the most accurate portrayal of the structure of the criterion space is provided by the interrelationships among the true scores. Estimating true score correlations by correcting for attenuation is a dangerous business which must be carefully done. The reliabilities that were used for Tables 5 and 6 are conservative in that they do not account for all the sources of error that might account for unreliability. For example, variability across testing occasions is not counted here but it might in fact serve to lower the correlations between pairs of variables (e.g., hands-on and knowledge tests).

Looking at the true score intercorrelations it seems reasonable to conclude that the hands-on measures and the knowledge tests designed to be parallel to them share a significant proportion of their variance. Also, the knowledge test designed to have a higher association with the job sample measure in fact does. Finally, the Army-wide rating measures of general soldier performance are by no means independent of the job sample measures but they have less in common than do the knowledge tests. One very large difference between tables 9 and 10 is in the lower correlations between the ratings and the other variables, particularly the ratings of specific task performance for 71L. However, administrative specialists tend to work more in isolation than other MOS and aren't observed as closely. It all seems to make reasonable sense.

Table 9

Intercorrelations among selected criterion measures for 11B.

Correlations corrected for attenuation are above the diagonal.

| | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Total score on all HO tasks | 1) | ( ) | 67 | 86 | 60 | 57 |
| Avg. of 15 HO task ratings (Sup. + Peer)[1] | 2) | 41 | ( ) | 44 | 39 | 73 |
| Total score on job knowledge KN test | 3) | 55 | 35 | ( ) | 80 | 31 |
| Total Score on training knowledge test | 4) | 40 | 33 | 70 | ( ) | 25 |
| AW-BARS - Overall Effectiveness (Sup. + Peers) | 5) | 32 | 51 | 23 | 19 | ( ) |

[1](Sup. + Peer) means that the corresponding correlations for supervisory ratings and for peer ratings were simply averaged.

Table 10

Intercorrelations among selected criterion measures for 71L.

Correlations corrected for attenuation are above the diagonal.

| | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Total score on all HO tasks | 1) | ( ) | 28 | 76 | 73 | 35 |
| Avg. of 15 HO task ratings (Sup. + Peer)[1] | 2) | 20 | ( ) | 10 | 29 | 51 |
| Total score on job knowledge KN test | 3) | 52 | 07 | ( ) | 82 | 22 |
| Total score on training knowledge test | 4) | 54 | 23 | 63 | ( ) | 27 |
| AW-BARS - Overall Effectiveness (Sup. + Peers) | 5) | 25 | 39 | 16 | 22 | ( ) |

[1] (Sup. + Peer) means that the corresponding correlations for superviory ratings and for peer ratings were simply averaged.

Participant and Administrator Reactions

In addition to the empirical results, the field tests provided considerable information about the logistical and administrative problems associated with collecting such a massive amount of information from each individual. The pooled feedback from both participants and adminstrators turned up fewer problems than most project members had expected. Such a project, organized in this fashion, can is fact be done. The first and probably most important hurdle is to gain the commitment of the people at the research site. Once the relevant personnel at the site truely believe that the project will start when the schedule says it will and genuinely accept their designated responsibilities then virtually all problems are solvable.

A major worry at the outset was the "motivation" of the participants themselves. It turned out to be far less of a problem than most people expected. While there were the inevitable exceptions, virtually all the participants seemed to take matters quite seriously and appeared to expend their best effort; even on the knowledge tests. A standard comment in that regard was, "well, we <u>are</u> supposed to know this stuff." The criterion measures that presented the most difficulty were the rating scales. Like everyone else, Army enlisted personnel do not like to make formal evaluations of other people, and frequently expressed discomfort at having to do so.

## Interpreting and Using the Field Test Results

The above results are only a bare summary of the complete data banks that were prepared for each MOS. Each data bank contained item and scale analyses, intercorrelations down to the scale and subscale level, and factor analyses of selected data sets. These data were carefully scrutinized by the previously described criterion measurement group. The group included the principal investigator for each of the criterion measures. Consequently, for each variable there was at least one committee member with a strong vested interest.

The other members of the committee consisted of the principal scientist for the project, the Army Research Institute contract monitors for the criterion development portions of the project, the Army Research Institute's chief scientist for the project, and one hapless individual who had to serve as chair (the second author of this paper) - ten people in all.

Again, the objectives of the group were to review the results of the field tests and to agree upon the specific revisions that were to be made in each criterion measure before the criterion array was declared the set of criterion measures that would be used for the concurrent validation. The mode of operation was for the principal investigator responsible for each criterion to review carefully the relevant field test data and propose the specific revisions, additions, or deletions that would maximize the reliability, acceptability, and construct validity of the job performance measures. A general discussion then followed, and continued until the investigator's proposal was accepted or a consensus was reached on what specific changes should in fact be made.

As a result of these discussions, the self ratings were dropped but both supervisor and peer ratings were retained, some hands-on tasks were

dropped and others were revised, both sets of knowledge tests were reduced in length by 20-30 percent, and the overall proportion of difficult items was reduced. In spite of the high intercorrelations among the rating scales, all the individual scales were retained. The primary reasons for retaining all scales was their apparent face validity to the participants and the participants' resistance to suggestions for eliminating or combining scales. On the notion that it is a relatively simple matter to sum scales later, none were eliminated. However, a number of changes were made in the scale directions and rater training procedures in an attempt to make them easier to use and to further increase their reliability.

The obvious disadvantage of the committee approach to data interpretation is the time involved. More than once the membership wished for a good dose of totalitarian power. On the positive side, all the major benefits of participative decision making seemed to manifest themselves. Everyone concerned always knew what was being done, crucial issues tended not to get lost, investigators could exercise veto power if the integrity of their product was being threatened, and considerable commitment seemed to have been generated. On balance, the time investment seemed worth it. In truth, on such a large multi-faceted project it probably is not possible for one "expert" to make these decision unilaterally. If the Project A model is used in the future with any frequency, applied psychologists must learn how to "manage" data interpretation as well as data collection.

## A Model of Job Performance

One of the major goals of the criterion component of Project A is to describe a comprehensive model of performance for entry level skilled jobs that makes sense conceptually and empirically, and which would be useful in future work on performance measurement. Whether or not the project can actually achieve such a goal must wait until the revised measures are used with larger samples. What we have currently are data on 100-150 people in each of nine jobs using pilot test versions of the instruments. However, in spite of being premature, we would like to make the following points about modeling the criterion space.

A basic point that should not generate argument is that job performance is multi-dimensional. There is not one attribute, one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is perhaps a bit more arguable to go on from there and assert that job performance is a construct (which implies a "theory" of performance), and is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization. For example, a manager could make contributions to organizational goals by working out congruent short term goals for his subordinates, and thereby guiding them in the right direction, or by praising them for a job well done, and thereby increasing subsequent effort levels. Each of these activities probably requires different knowledges and skills which are in turn most likely a function of different abilities. Consequently, for any particular job, one fundamental task of performance measurement is to describe the basic factors that comprise performance. That is, how many such factors are there and what is their basic nature?

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 is a value judgment on the organization's part. The determination of the specific combinational rules (e.g., simple sum, weighted sum, non linear combination) that best reflect what the organization is trying to accomplish is in large measure a research question. In sum, it makes sense to assert that performance in a particular job is made up of several relatively independent components and then ask how each component relates to some continuum of overall utility. It is quite possible for people with quite different strengths and weaknesses on the performance factors to have very similar overall utility for the organization.

If performance is characterized in the above manner, then a more formal way to model performance is to think in terms of its latent structure. The usual common factor model of the latent structure is open to criticism because all of the criterion (i.e. performance) measures may not be at the same level of explanation or they may be so qualitatively different that putting them into the same correlation matrix doesn't seem appropriate. For example, combining the dichotomous variable stay vs. leave (voluntarily) with a hands-on job performance test score seems like a strange thing to do. Also, two criteria may not be functionally independent. One may be a cause of the other. The situation can be even further complicated if the causal variable (e.g., a knowledge test of training content) is a "purer" (i.e. more

construct valid) measure of the latent variable of real interest than is the final variable in the causal chain (e.g., a knowledge test of job content).

Conceptual gymnastics such as the above have led some people to propose structural equation modeling as a way to portray more meaningfully the criterion space and the associated predictor space (e.g., Bentler, 1980; James, Muliak, & Brett, 1982).

From this perspective, the aims of criterion analysis are to use all available evidence, theory, and professional judgment to A) identify the variables that are necessary and sufficient to explain the phenomena of interest and B) specify the nature of the relationships between pairs of variables in terms of whether they are 1) correlated because one is a cause of another, 2) correlated because both are manifestations of the same latent property, or 3) are independent. The more explicitly the causal directions and the predicted magnitude of the associations can be specified the greater the potential power of the model. That is, it more clearly outlines the kinds of data to be collected and the kinds of analyses to be done; and it provides a much more explicit framework within which to interpret empirical results.

Within the structural equation framework there are two general kinds of models, one dealing with manifest variables (operational measures) and one with latent variables (constructs). The most thorough portrayal of a domain involves both. Certainly we have assumed that it does in Project A. The proposal and research plans have talked explicitly about criterion constructs and criterion measures. What we really want to model, in terms of identifying the necessary and sufficient variables and their causal interrelationships, are the more "fundamental" underlying constructs. What we in fact will have are operational measures that represent the constructs (hopefully).

A textbook illustration of a latent structural model and its associated operational measures is shown in the attached figure from James, Muliak, & Brett (1982), p. 121.

- - - - Figure 2 About Here - - - -

A few points, some general - some specific, should be made about such a picture.

First, it is true that we simply know a lot more about predictor constructs than we do about job performance constructs. There are volumes of research on the former, and almost none on the latter. For personnel psychologists it is almost second nature to talk about predictors in terms of constructs. Investigation of job performance criterion constructs seems limited to those few studies dealing with synthetic validity and those using the critical incident format to develop performance factors. Relative to the latter, the jobs receiving the most attention have been managers, nurses, firefighters, police officers, and perhaps college professors (cf Landy & Farr, 1983).

Second, the usual textbook illustration of a latent structural equation model shows each latent variable being represented by one or more manifest operational measures. However, in our situation, just as it is easy to think of examples where a predictor test score could be a function of more than one latent variable (e.g., A computerized two-hand tracking apparatus could be a function of several latent psychomotor "factors"), the same will be true of criterion measures.

Third, we would be hard-pressed to defend placing the criterion variables on some continuum from immediate, to intermediate, to more ultimate as a means for portraying their relative importance or functional inter-

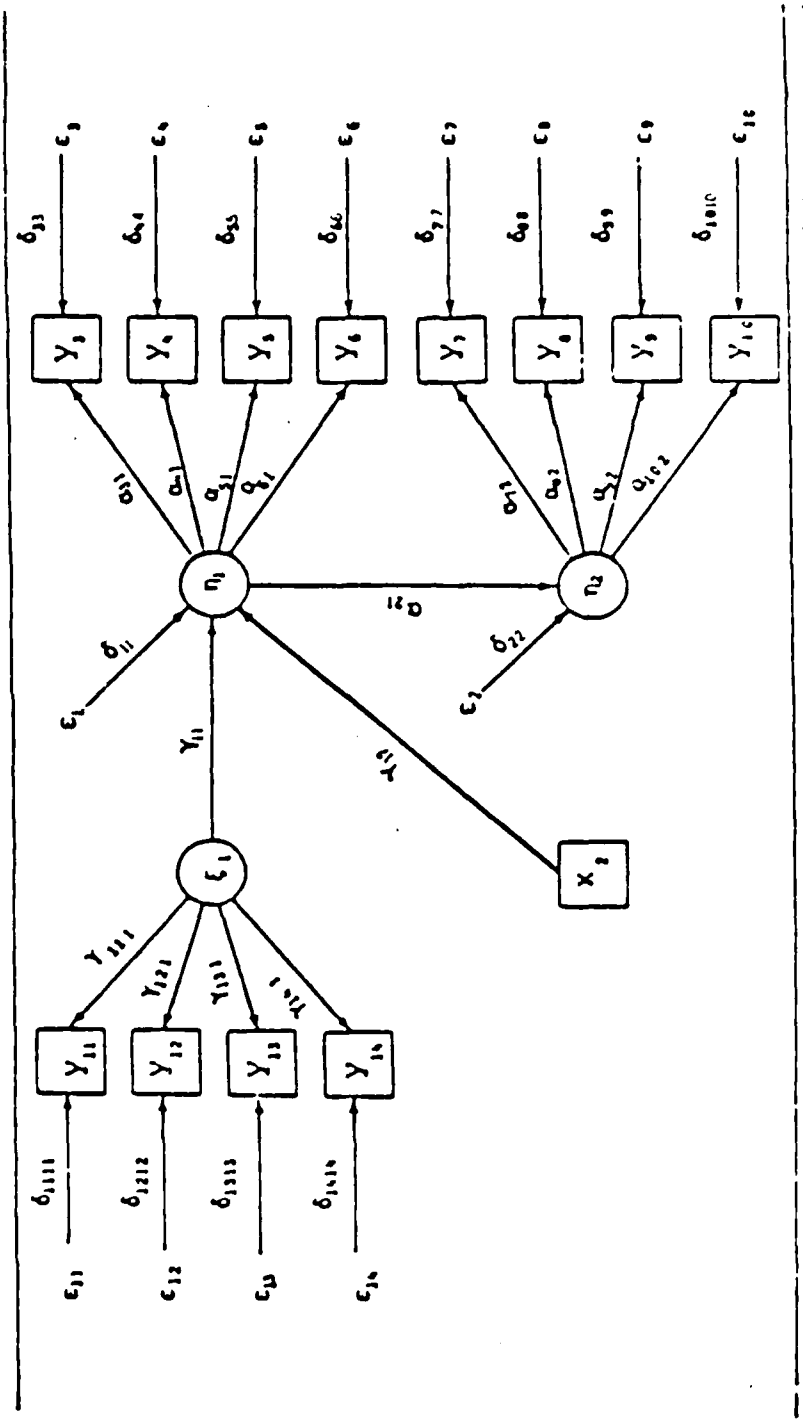


Figure 4.9 A structural equation model in which two exogenous variables, $\xi_1$ and $\xi_2$, are causes of a latent variable $\eta_1$, which in turn is a cause of latent endogenous variable $\eta_2$. Each of the latent variables in the model is represented by four manifest variables.

Figure 2

relationships. For example, although there are good reasons for developing hands-on performance measures, is there any among us who would be willing to defend hands-on performance scores as the "most ultimate" measure? We hope not.

Fourth, it is also true that the development of the criterion variables was guided to a certain degree by what we thought were the constructs/factors making up enlisted performance. The work was also guided by the desire to cover as many bases as possible relative to the population of criterion measures that it is possible to collect. That is, because we know so little about the latent structure of job performance, we used every bit of technology we had.

Fifth, because we have such a mixture of a large number of variables, specifying a formal structural model that can indeed be tested with LISREL V would be difficult. However, as a means for focusing discussions and arguments about criterion relevance, criterion combination, criterion equivalence, etc., attempting to construct such a model should be very useful.

Sixth, a conjoint model would most likely come closest to revealing the "truth" about how the criterion measures should be combined for specific decision making purposes. That is, it is probably the case that the score or value assigned to a particular level on one variable is a function of the individual's standing on another variable (e.g., if an individual if going to attrit, it doesn't matter what his or her hands-on performance score is). Conversely, if an individual is a poor performer on a hands-on measure, we may want him or her to leave but if the individual is a good performer, leaving the Army is bad. (It gets complicated very quickly). One problem with applying a conjoint model is that actually estimating the parameter values rapidly becomes a very complicated data collection task.

The exercise

The next step will be to try our hand at developing a tentative structural model. This will involve a lot of arguing about things like:

1) What is the best portrayal of the latent variables in the criterion space?

2) What is the best estimate of how these latent variables are interrelated? Specifically: On the basis of the hypothesized latent structure -

   a) For each pair of manifest variables, should there be a causal relation via common latent variable, or no relation?

   b) Should a particular non zero relationship be positive or negative, linear, or non linear?

   c) If non linear, why?

We don't have very good answers to such questions at present, but as a first attempt at portraying the latent structure suppose we suggest that the enlisted performance domain is made up of the following general factors.

1) Maintaining and upgrading current job knowledge (including common tasks). A legitimate question here might be why the mere possession of job knowledge should be a factor in the performance domain. However, if a major goal of the Army is to be ready to enter a conflict on short notice, then possessing a high degree of current knowledge is performance. Having the proper information and being able to use it (factor 1) are not the same thing. However, neither are they independent. Consequently, our model must stipulate that these first two factors are significantly correlated and the relationship stems

both from sharing common requirements (e.g., general cognitive ability) and because factor two is in part a "cause" of performance differences on factor one.

2) Technical proficiency on the primary job tasks. This factor refers to being able to perform on the technical content, be it complicated or simple. Technical is defined broadly but not so broadly as to include leadership or other interpersonal interaction task requirements. Within this construct the content of the tasks may vary considerably and rely on very different abilities (e.g., playing a musical instrument vs. repairing a truck generator). For most jobs it might also be possible to think of two such general factors, the execution of "standard" procedures and troubleshooting special problems.

3) Exhibiting peer leadership and support. It is often the case that enlisted personnel have the opportunity to teach, support, or provide leadership for their peers. This factor refers to the frequency and proficiency with which people do that when the occasion arises. It would also be reasonable to think of this factor as composed of the four subgeneral factors that have been found in leadership research (e.g., Bowers & Seashore, 1966), goal setting, facilitating goal attainment, one-on-one individual support, and facilitating group morale.

4) Demonstrating commitment to Army regulations and traditions. Performance on this factor refers to maintaining living quarters equipment, and maintaining a high level of physical fitness and appropriate military appearance. This factor is perhaps a bit more tenuous than the others. Defining it this way assumes

that all of the different elements will covary to a high degree.

5) Continuing to perform under adverse conditions. This factor would share many components in common with the previous three and thus should not be orthogonal. However the act of carrying out job assignments when wet, tired, or in danger is viewed as a very important and distinct aspect of performance.

6) Avoiding serious disciplinary problems. Incurring disciplinary actions because of problems with drugs, alcohol, neglect of duty, or serious interpersonal conflict represent a great cost to the Army. Successfully avoiding these costs is viewed as an important factor in overall performance.

Standing in a direct causal relation to the performance factors are knowledges and skills learned during training, abilities and other individual characteristics present at the time of hire, and the choice to perform, which is supposedly under motivational control. For our purposes here, the causal latent variables of most concern are the knowledges, skills, and motivational predispositions acquired during training. Consequently, we might posit that there are three major training performance factors in the latent structure.

1) Hands-on task proficiency.

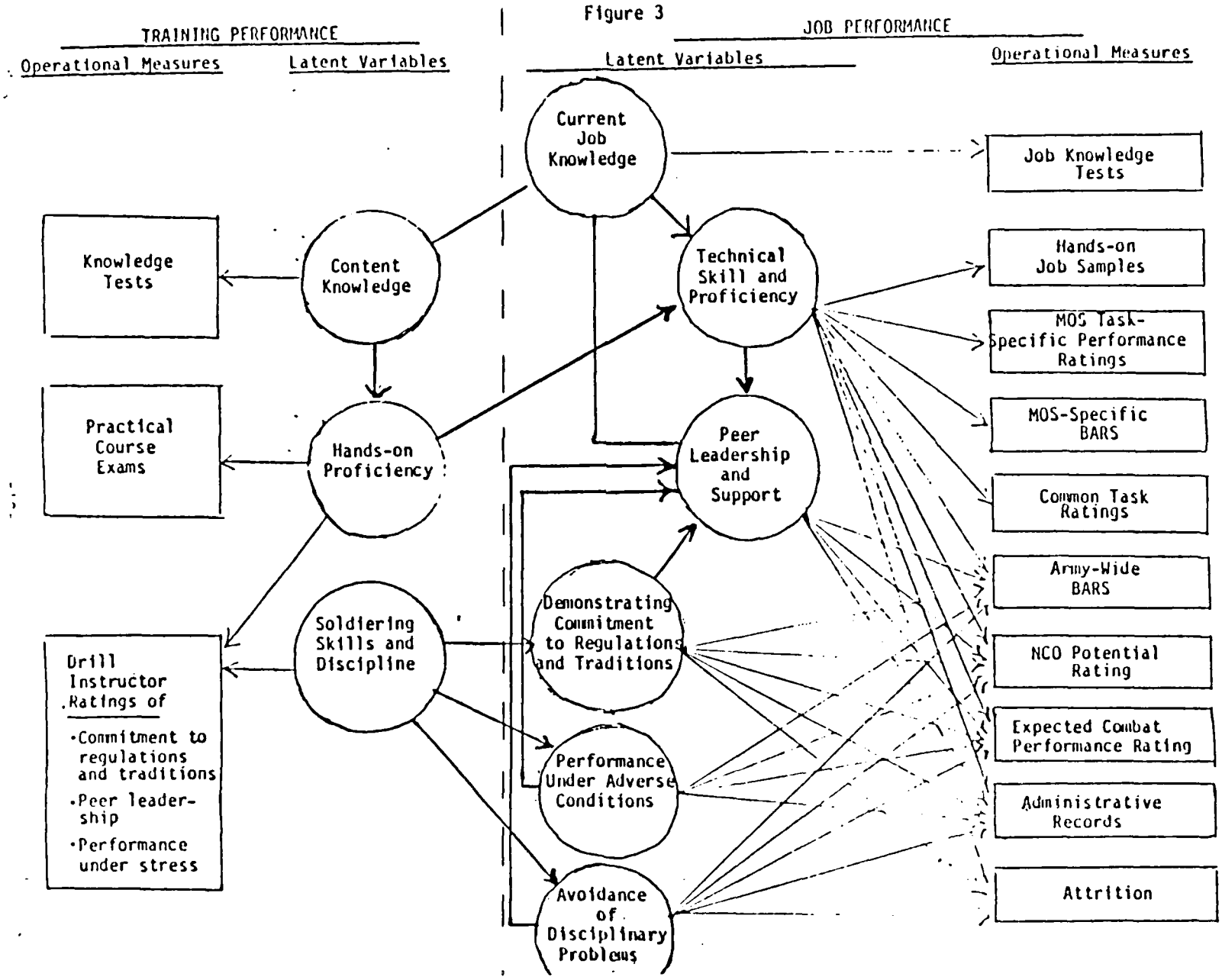
2) General job knowledge.

3) Exhibiting good soldiering skills and discipline.

A very rough schematic that portrays these latent variables and also lists the observable measures of them that we have available in Project A is shown in Figure 3. The arrows between latent variables and operational measures indicate an expected correlation. The expected size of the correlation is not indicated. Arrows between latent variables indicate a hypothesized causal relation.

- - - - Figure 3 About Here - - - -

Figure 3


TRAINING PERFORMANCE
Operational Measures
Latent Variables
JOB PERFORMANCE
Latent Variables
Operational Measures
Current Job Knowledge
Job Knowledge Tests
Knowledge Tests
Content Knowledge
Technical Skill and Proficiency
Hands-on Job Samples
MOS Task-Specific Performance Ratings
Practical Course Exams
Hands-on Proficiency
Peer Leadership and Support
MOS-Specific BARS
Common Task Ratings
Army-Wide BARS
Soldiering Skills and Discipline
Demonstrating Commitment to Regulations and Traditions
NCO Potential Rating
Drill Instructor Ratings of
•Commitment to regulations and traditions
•Peer leadership
•Performance under stress
Performance Under Adverse Conditions
Expected Combat Performance Rating
Administrative Records
Avoidance of Disciplinary Problems
Attrition

some issues and problems with which we must deal.

First, the principal data upon which the list of latent constructs is based are the results of the critical incident workshops conducted during the development of the behaviorally anchored rating scales. We have not yet had the opportunity to examine the factor structure of the hands-on measures or knowledge tests, or even to look comprehensively at the factor construct of all the rating scales. These analyses really must wait until larger sample sizes are available with the revised measures. Such data are currently being collected as part of Project A's concurrent validation sample where N's will be 500-700 for each MOS.

Second, the manifest job performance variables are by no means "pure" measures of the latent constructs. For example, factor two would seem to underly virtually all of the observable measures. By contrast the "avoidance of disciplinary problems" should influence only some of the army-wide BARS scales, NCO potential, attrition, and perhaps expected combat performance. However in general, most of the observable variables are probably multiply determined.

Third, the above reasoning suggests that if the operational criteria share so many common determinants they probably should not receive grossly differential weights when combined into composites for the purpose of test validation.

Fourth, differential prediction of job performance across jobs must come from different requirements for success on factors one and two (e.g., psychomotor abilities vs. verbal ability). To a certain extent it could also result from a greater weight being given in some MOS to peer leadership and performance under adverse conditions.

Fifth, limiting measures of training success to paper-and-pencil tests of

knowledges mastered is probably not sufficient. To more completely determine the relationship of training performance to job performance additional measures would be required.

Sixth, the causal relations among: the individual differences present at the time of entry; the latent variables making up training performance; the latent variables that constitute job performance; and the operational criterion measures, can be described with brilliant understatement by saying that they are complex. As part of that complexity it seems reasonable to assert that:

- Among the latent variables describing training performance, hands-on proficiency and content knowledge are more highly related to each other than either is to soldiering skills and discipline. Further, content knowledge stands in at least a partial causal relation to hands-on proficiency.

- Among the latent variables describing job performance, job knowledge would seem to come first in the causal chain since it at least partially determines technical proficiency. However, both these factors most likely cause at least some of the individual differences in peer leadership performance. A causal relation between technical proficiency and either commitment to regulations/traditions or avoiding disciplinary problems does not seem so likely. However, some may wonder whether or not commitment to regulations/traditions and avoidance of disciplinary problems are bipolar.

- If the first two factors were measured with high construct validity then factor one (current job knowledge) should have a direct effect only on job knowledge tests. Job knowledge should create

differences on other operational measures only through its influence on technical proficiency. Consequently, if technical proficiency could be held constant the observed correlations between job knowledge tests and all other variables should be reduced to zero.

- Since peer leadership and support was given a broad definition (in terms of leadership theory), greater knowledge, higher technical skill, higher commitment, demonstrated performance under stress, and an exemplary record would all "cause" an individual to exhibit more effective peer leadership.

- As somewhat of a contrast, performance under adverse conditions is conceptualized as a dispositional variable. Consequently it would be under motivational control and not a function of knowledge or ability.

## A Final Comment

The above comments are still highly speculative. Such a model of performance will go through many iterations before the project is finished. Even this brief discussion illustrates that when the task force meets again to consider the analysis and interpretation of the data from the concurrent validation itself there will be lots to argue about. The pros and cons of scientific dictatorship first participatory democracy will be hotly debated. We hope that at least some of us will survive, and the project as well.

## REFERENCES


Bentler, P. M. Multivariate Analysis. In M. Rosenzweig & L. Porter (Eds.) Annual Review of Psychology, Vol. 30. Palo Alto: Annual Reviews, Inc.,

Bowers, D. G., & Seashore, S. E. Predicting organizational effectiveness with a four factor theory of leadership. Administrative Science Quarterly, 1966, 11, 238-263.

Eaton, N. K., & Goer, M. H. (Eds.). "Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Technical Appendix to the Annual Report," Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences, ARI Research Note 83-37. (1983)

James, L. R., Muliak, S. A., & Brett, J. M. Causal analysis: Assumptions, models, and data. Beverly Hills, CA: Sage, 1982.

Landy, F. J., & Farr, J. L. The measurement of work performance: Methods, theory, and application. New York: Academic Press, 1983.

Sackett, P. R., & Dreher, G. F. Constructs and assessment center dimension: Some troubleshooting empirical findings. Journal of Applied Psychology, 1982, 67, 401-410.

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

# MEASUREMENT OF ENTRY-LEVEL JOB PERFORMANCE

by

Newell K. Eaton

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

August 1985

Paper presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

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

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

# MEASUREMENT OF ENTRY-LEVEL JOB PERFORMANCE

Newell K. Eaton[1]

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

The purpose of this paper is to introduce the research project from which the data are drawn for the following papers on combining knowledge and hands-on measures, performance ratings, and criterion reduction. I hope it will also provide an overview of the criterion measurement strategy so as to place those papers in the context of the larger effort. I would also like to add that much of the criterion measures research discussed today is part of the Army's contribution to the Joint Services criterion measurement project described by Bert Green and his colleagues in a symposium here at APA Saturday.

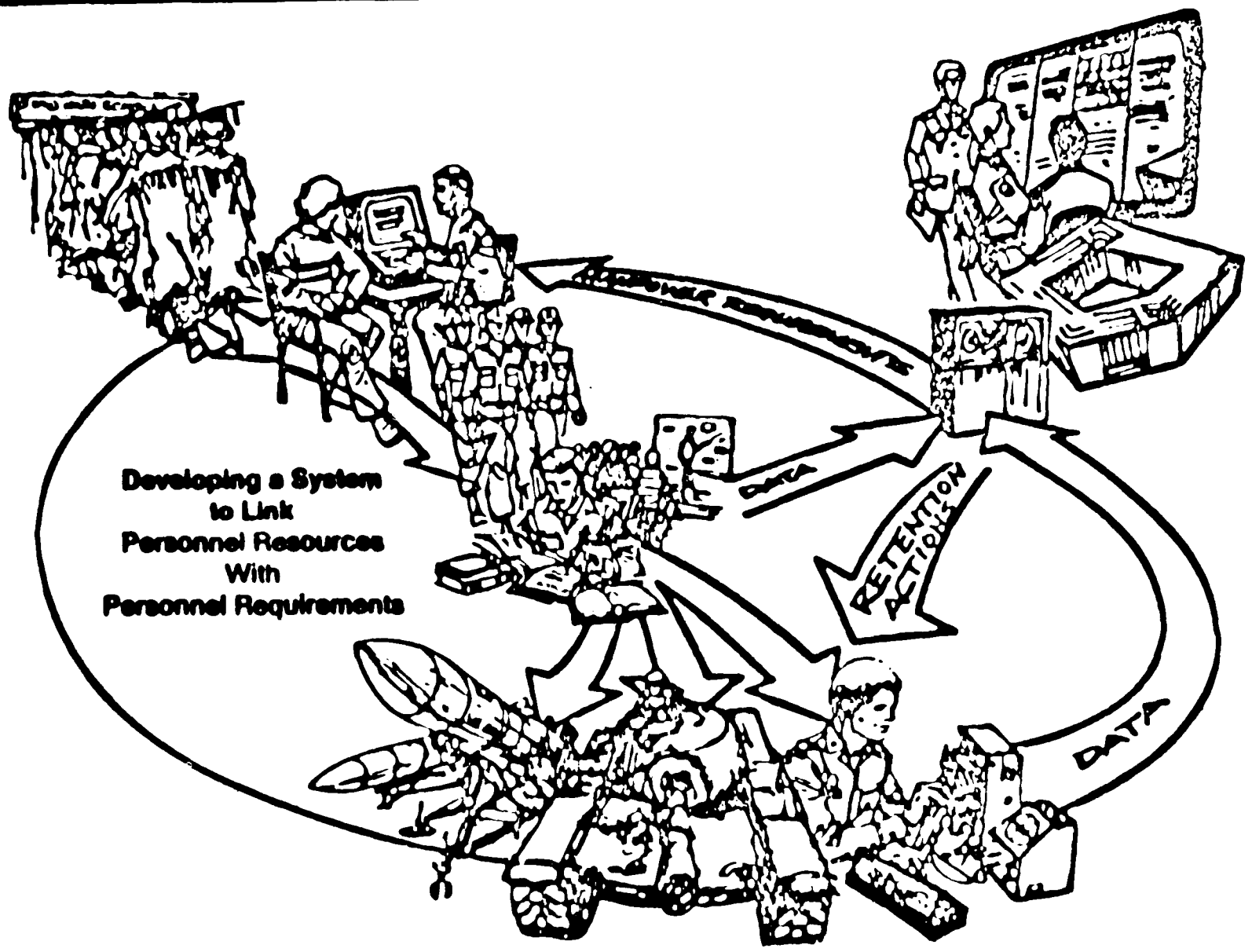
---

[1]Portions of this paper have been adapted from Eaton, N. K., Goer, M. H., and Zook, L. M. (1984) Introduction to current Army selection and classification research. In N. K . Eaton, M. H. Goer, and L. M. Zook (Eds.) Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 Fiscal Year (Technical Report 660). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, and Eaton, N. K., Hanser, L. M., and Shields, J. S. (In Press) Validating selection tests against job performance. In J. Zeidner (Ed.) Human productivity enhancement. New York: Praeger.

In 1982 the Army began a comprehensive 9-year selection, classification and assignment research program. The goal is a computerized personnel allocation system to match available personnel resources with Army manpower requirements, based on biographical, psychological, and performance measures and a firm quantification of their interrelationships. "Project A" will develop, for first- and second-tour soldiers, new predictor tests and composites, performance measures, composites, and utility values, and an empirical description of their intercorrelations. These, along with supply and demand forecasts, will be the basis for the concurrent development, by "Project B," of the computerized allocation system. The major features of the program are illustrated in Figure 1. These major features are the new predictor tests, their empirical linkage to both training and job performance, the use of all these data in reenlistment decisions, the allocation system for enlistment and reenlistment based on these data, and the determination of manpower and personnel requirements based on these data.

The research program is designed to facilitate the management of the U.S. Army enlisted force. This is one of the most complex personnel management tasks in the world. Each year over 400,000 people apply for 135,000 first-tour positions in over 250 military occupational specialties. Over 80,000 soldiers reenlist in about 350 different occupations. Typically, an individual is guaranteed specific occupational training at the time he or she signs an enlistment contract, and a specific occupation upon reenlistment. Enlistment can be up to one year prior to entering the Army. The decision to select the individual

# THE ARMY'S PERSONNEL SYSTEM



for service or reenlistment and to allocate an occupation must be made to meet the needs of the individual as well as the near-term requirements and long-range objectives of the Army.

Of course, the Army is not now without tools for making such decisions. Standards are in place for initial selection and classification. They have been shown to be valid for training performance and job knowledge in many occupations. A system does exist for occupation allocation in enlistment and reenlistment. With the accomplishment of Projects A and B, however, the Army's personnel system will be far superior to existing systems, benefiting both individual soldiers and the productivity of the Army.

A major effort to develop new predictor and performance measures is being conducted to expand the dimensionality and accuracy of measurement of the respective predictor and criterion spaces. At this time there appears to be a heavy general-ability loading in both the paper-and-pencil Armed Services Vocational Aptitude Battery used for selection and classification, and the current paper-and pencil job knowledge tests, called Skill Qualification Tests. The research described here is designed to provide measures that more completely encompass the full range of potential performance prediction, and to provide criterion measures that more adequately represent actual job performance. Together, these should enable the Army to make the most valid performance predictions. An improved personnel management system, based on a variety of better predictor and performance measures, from an appropriate

sampling of representative Army occupations, is illustrated in Figure 2. In each occupation the most valid composite from a full range of predictors will be used as selection and classification factors to provide the best person-job match for overall soldier performance.

## Research Design

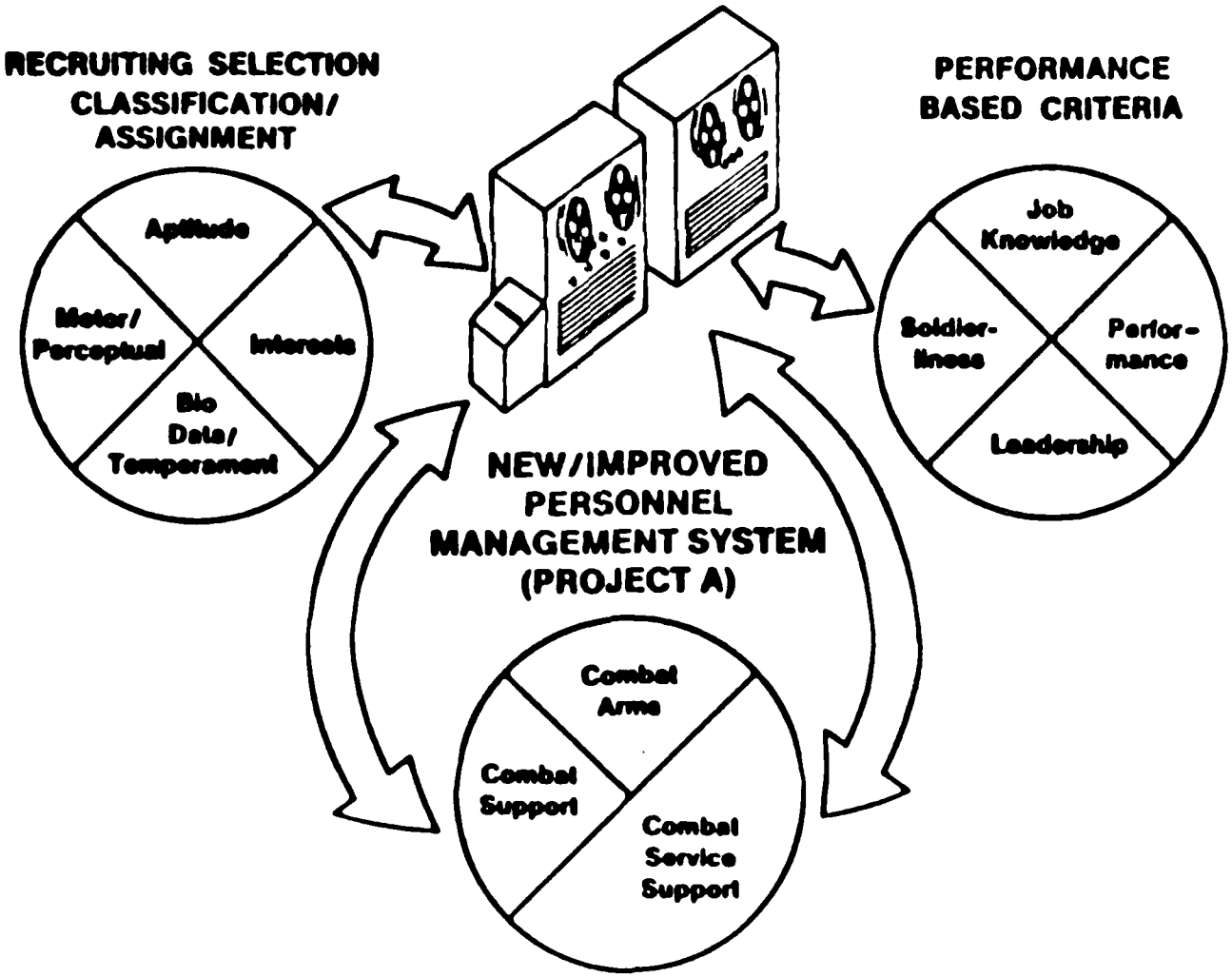
The Project A research design is shown in Figure 3. A key feature of the design is its iterative nature. Data are being collected in three iterations to provide for timely and responsive results during the course of the effort, as well as to correct for errors and to take advantage of opportunities.

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

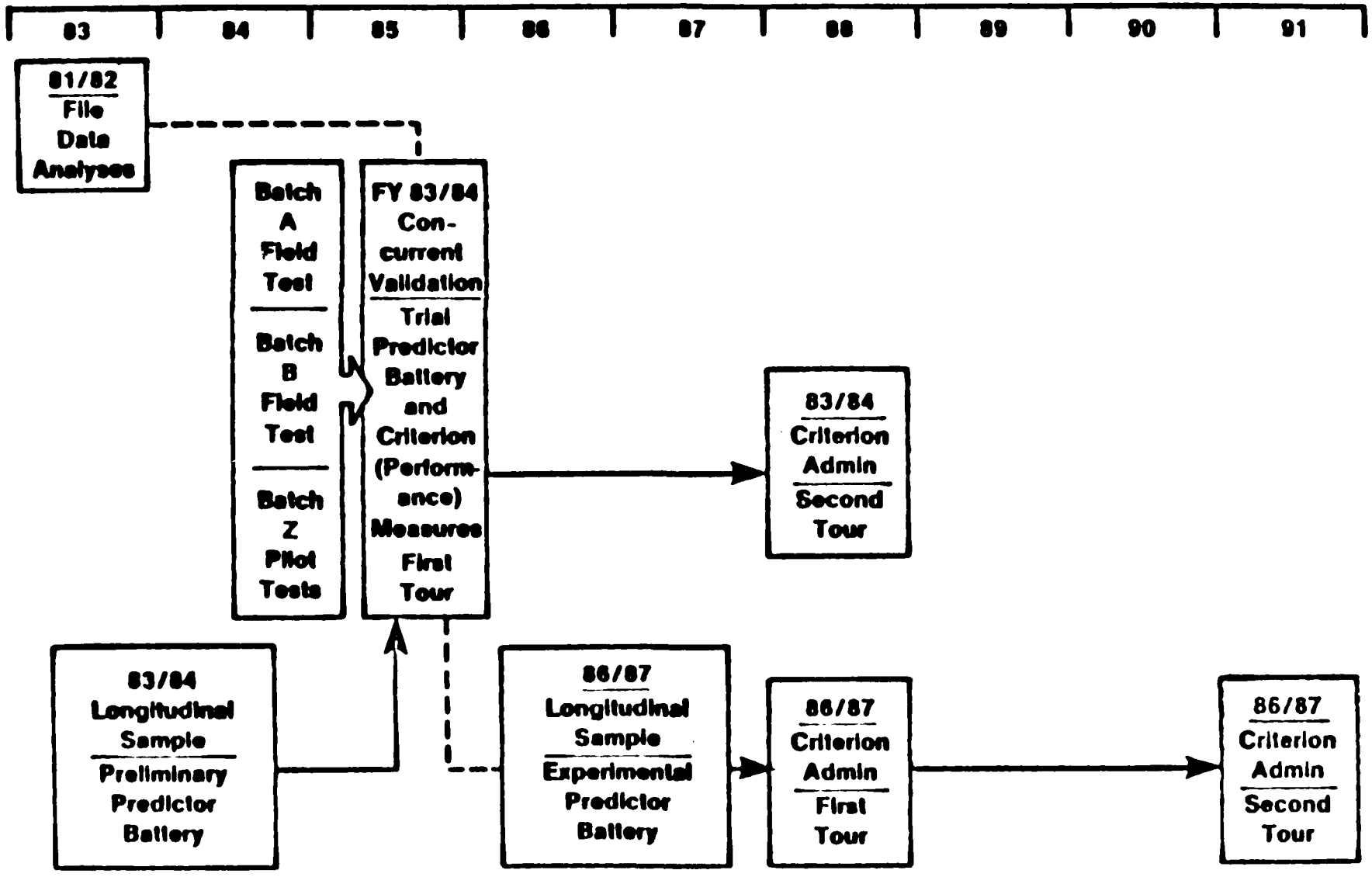
In the second iteration about 12,000 first-tour soldiers are participating in a concurrent validation effort this summer. About 600 soldiers in 19 representative occupations are being tested. The testing this summer is based on an extensive field testing effort conducted last summer, fall, and winter. A revised test battery, including computer-administered perceptual and psychomotor predictor instruments, is being concurrently administered with a set of job-specific and general performance indices and rating measures. About a hundred soldiers in each occupation will be retested after three years, during their second Army tour.

Figure 2

# SYSTEM OBJECTIVES



# THE RESEARCH FLOW



The 19 occupations chosen for testing comprise a specially selected representative sample of the Army's 250 entry-level occupations. Occupation selection was based on an initial clustering of occupations, derived from rated similarities in content. These 19 occupations chosen account for about 45 percent of Army accessions. In most of the occupations sample sizes are sufficient to evaluate empirically both race and sex fairness.

In the third iteration all of the measures, refined by the experiences of the first and second iteration, will be collected sequentially in a predictive validity design. About 50,000 soldiers across about 20 occupations will be included in the fiscal year 1986 and 1987 predictor battery administration. After losses from all factors, about 3,500 will be included in second-tour performance measurement in fiscal year 1991.

The Criterion Space

The design of the research has been driven by the desire to measure job performance comprehensively and to assess the utility of differences in individual job performance to the organization. Because of the dimensionality of job performance, many different performance measures are not only possible, but desirable. The problem is to identify the fundamental factors contributing to successful performance in a specific job and then to develop appropriate measures. The selection of the appropriate measures has been a matter of judgment and analysis.

The first step in developing the performance measures was an analysis of each occupation. An extensive task inventory for each of the

19 key occupations was developed, based on Soldier's Manuals, and official occupational surveys, and also from numerous subject matter experts. Efforts were made to understand the behavioral aspects of each task, to standardize the generality of task descriptions, and to determine the variability of performance, importance, and frequency for each task. In addition, critical incident workshops were conducted with non-commissioned officers and commissioned officers. They generated examples of effective and ineffective performance as well as those general aspects of soldier effectiveness that contribute to organizational effectiveness, such as following orders and regulations. As a consequence the target criterion space went beyond specific job performance to include aspects of socialization and commitment to the organization. Using the critical incidents generated in these workshops, job-specific and Army-wide performance dimensions were identified and defined.

Criterion Measures

Criterion measures consist of hands-on tests, paper-and-pencil based job knowledge tests, and both peer and supervisory ratings. The final selection of tasks for hands-on and job knowledge testing was based on task importance, task difficulty and intertask similarity judgments from subject matter experts. Tasks were clustered based on intertask similarity. Tasks were selected from the clusters on the basis of importance and difficulty to represent each cluster.

On this basis, 30 tasks were selected for each occupation. Procedural knowledge, paper-and-pencil based measures, consisting of 3-16 items, were developed for each of these 30 tasks. One of the 30 tasks selected for mechanics is "troubleshoot electrical system." The first of the 10 items tested on this task is shown in Figure 4. It illustrates the extensive use of drawings in these paper and pencil measures of task knowledge, and the procedural nature of the items.

For a subset of 15 of the 30 tasks chosen for testing, hands-on measures were also developed. Tasks were chosen for hands-on measurement if they were judged to require a high level of physical skill, a series of prescribed steps, and speed of performance. The scoresheet for the hands-on test of the mechanic's task "troubleshoot electrical system" is shown in Figure 5. The hands-on items are scored "go" or "no go." The first action to be performed on this task is to check the alternator drive belt tension. This was also the first item in the procedural knowledge test of this task shown in the previous figure. In almost every case, the hands-on tests were parallel in item content with the procedural knowledge tests of the same task.

A simple 7-point rating scale was also developed for each of the 15 tasks selected for hands-on measurement for each occupation. Both supervisors and peers provided the ratings.

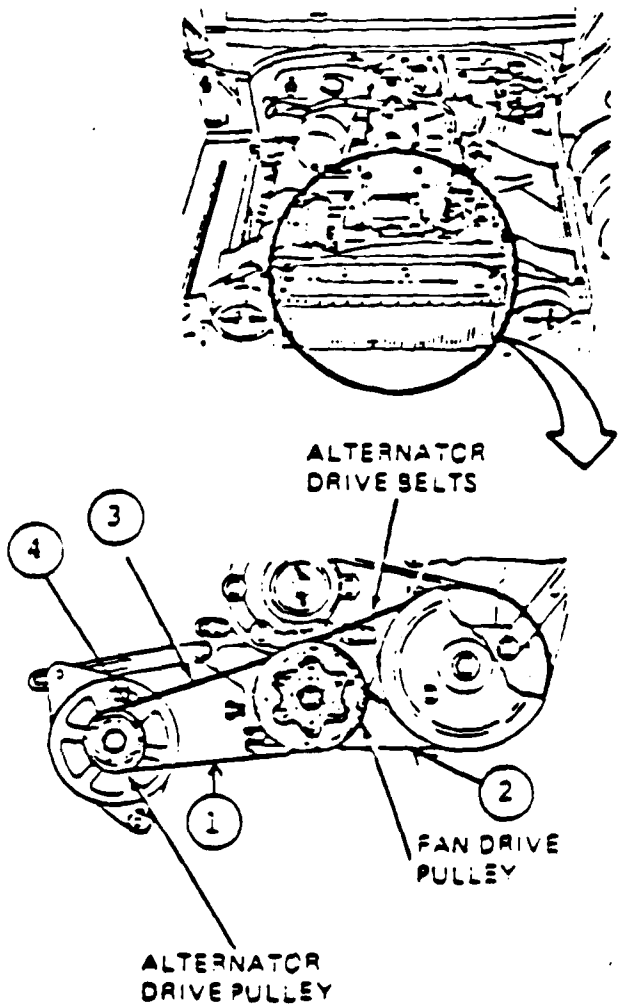
Figure 6 helps to summarize the task-based testing. Of the 30 tasks chosen for procedural knowledge testing in each occupation, 15

You must troubleshoot the electrical system of an M809 series vehicle.

1. You are checking the alternator drive belt tension. Where on Figure 1 below should you check the tension?

   A. 1
   B. 2
   *C. 3
   D. 4

Figure 4



SCORESHEET

Scorer: _____ Soldier: _____

Date: _____ SSN: _____

Note to Scorer: Tell the soldier: "YOU MUST TROUBLESHOOT THE ELECTRICAL SYSTEM ON THIS TRUCK. FIRST YOU SHOULD CHECK THE ALTERNATOR DRIVEBELT TENSION."

| PERFORMANCE MEASURES | GO | NO-GO |
|---|---|---|
| 1. Pushed down on one alternator drive belt midway between alternator drive pulley and fan drive pulley. Found tension correct. | ____ | ____ |
| 2. Pushed down on other alternator drive belt midway between alternator drive pulley and fan drive pulley. Found tension correct. | ____ | ____ |

Note to Scorer: Tell the soldier: "NOW YOU SHOULD TEST THE ALTERNATOR OUTPUT VOLTAGE"

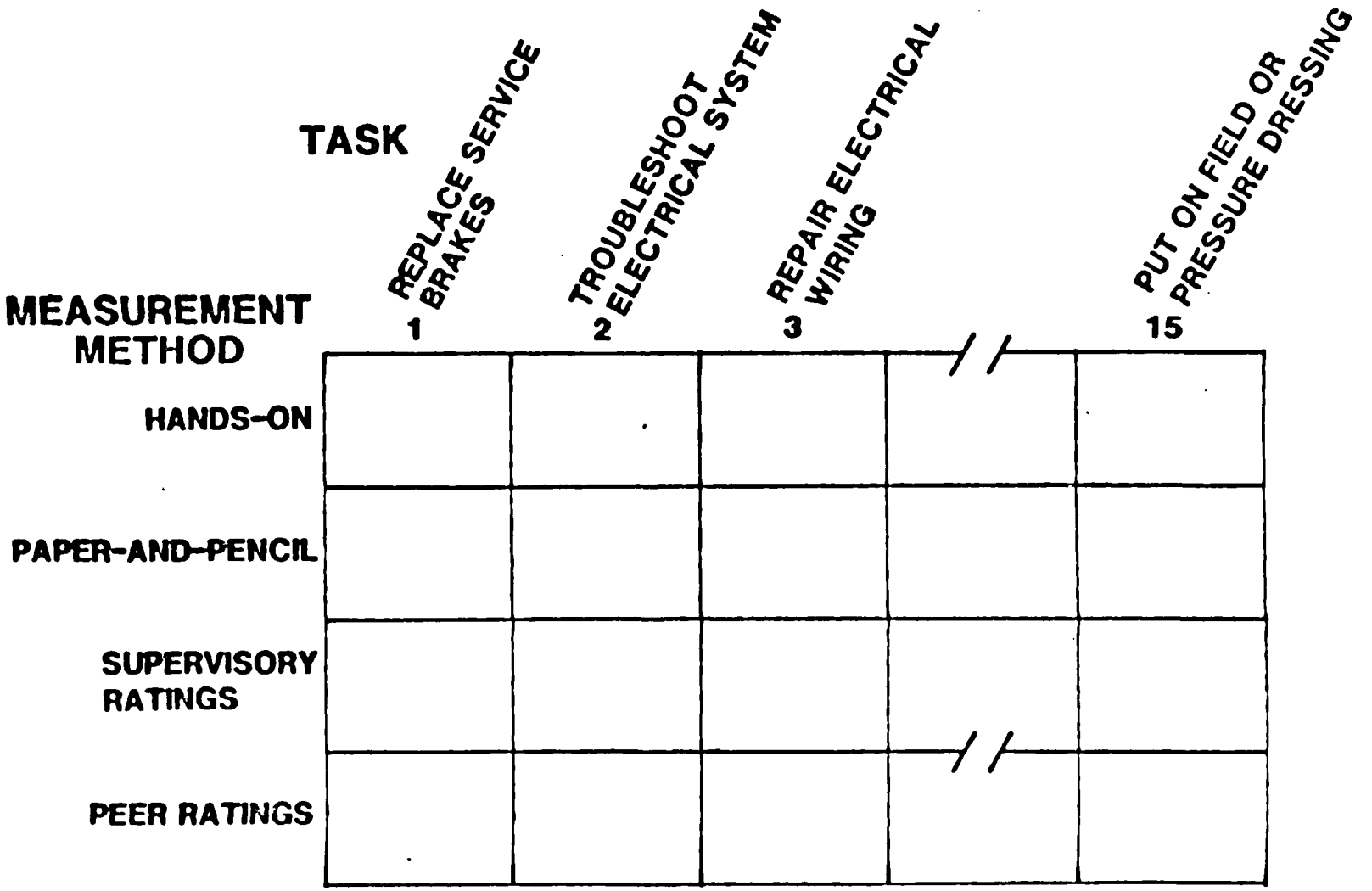
| PERFORMANCE MEASURES | GO | NO-GO |
|---|---|---|
| 3. Removed battery ground cable from battery. | ____ | ____ |
| 4. Removed alternator terminal cover. | ____ | ____ |
| 5. Reconnected battery ground. | ____ | ____ |

Note to Scorer: Stop the soldier now and say, "NOW INSTEAD OF ACTUALLY PERFORMING THE NEXT STEPS, I WANT YOU TO TELL ME WHAT YOU WOULD DO TO FINISH THE TASK. YOU CAN WALK THROUGH IT IF YOU WANT WHILE YOU TELL ME."

Figure 5

Figure 6

# MECHANIC'S TASKS SELECTED FOR MULTI-METHOD TESTING



were also evaluated through hands-on testing, and both peer and supervisory ratings. Some of these 15 tasks for mechanics are shown in the illustration.

This design is expected to allow a more complete evaluation of task performance as well as a better understanding of the types of tasks for which the different testing modalities are most appropriate.

More Rating Scales

Analyses of data from the behavioral workshops formed the basis for development of another set of performance measures. Eleven behaviorally-anchored rating scales were developed for application to soldiers regardless of occupation. In addition, six to nine behaviorally-anchored occupation-specific ratings were developed for each occupation. Last scales were developed to rate overall performance, individual potential, and performance on 14 Army tasks common to all occupations. Our goal was to obtain ratings on each soldier, with each rating scale, from two supervisors and four peers.

The Relationship between Task Clusters, Tasks Selected for Testing, and the BARS

The clusters of tasks for mechanics are shown in Figure 7. In the column of numbers on the left side of the illustration are shown the distribution of the 15 tasks specific to mechanics that were chosen for hands-on, procedural knowledge, and rating measurement. The center column shows the distribution of the 15 additional tasks from the cluster chosen for procedural knowledge testing only. The column on the right illustrates the occupation-specific rating dimensions measured by

Figure 7

Task Clusters and Rating Scale Dimensions for Mechanic (MOS 63B)

| Task Cluster | Multi-method Testing[1] | Knowledge-only Testing[2] |
|---|---|---|
| Routine Maintenance | 1 | 2 |
| Disabled Vehicles | 0 | 2 |
| Brakes | 3 | 1 |
| Carburetor, Radiator | 2 | 2 |
| Clutch, Powertrain | 1 | 1 |
| Electrical System | 2 | 1 |
| Steering, Drive Components | 1 | 1 |
| Personal Weapons | 1 | 1 |
| Individual Tactics | 0 | 2 |
| First Aid | 1 | 1 |
| Chemical/Biological Hazards | 1 | 1 |
| General Soldier Skills | 2 | 0 |
| | 15 | 15 |

| Occupation-Specific Rating Dimensions |
|---|
| Inspecting/Testing |
| Troubleshooting |
| Routine Maintenance |
| Repair |
| Technical Documents |
| Tools and Test Equipment |
| Vehicle Operation |
| Safety |
| Administration |
| Planning/Organizing Job |
| Vehicle Recovery |
| Overall Performance |

[1]Number of tasks tested by Hands-on, Knowledge, and Rating methods.

[2]Number of additional tasks tested by Knowledge test only.

BARS. Comparison of the task clusters on the left with the rating dimensions on the right illustrates the more general nature of the rating dimensions. The task clusters are primarily content based, and tasks within the clusters reflect actions or procedures related to that content, such as troubleshooting, repairing, or maintaining the electrical system. The rating dimensions are more process based, and cross task clusters. For example, the BARS troubleshooting dimension is specific to 3 hands-on tasks, and 4 additional knowledge only tasks, in five of the task clusters.

Multiple Choice Job Knowledge Testing

For each occupation a traditional multiple-choice job knowledge test was also developed. Each contains between 125 and 250 items per occupation. These were developed as training criterion measures, based on item budgets designed to reflect the incidence of performance of tasks within duty areas. While designed as training measures, more than 95% of the items are appropriate as both end-of-training and first tour measures, based on the similarily of content pertinent to these stages of the soldier's career.

To complement the measures of performance described, additional scales were developed to evaluate the impact of environmental and leadership variables on performance, and to estimate possible performance in combat situations.

## Criterion Measurement by Occupation

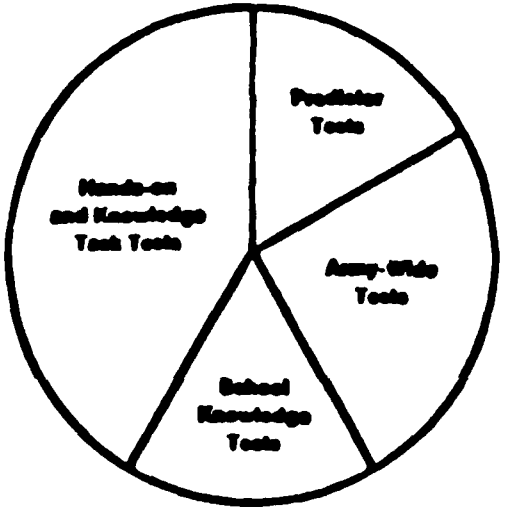
The complete set of performance measures was generated for 9 of the 19 occupations. For the remaining 10 occupations a partial set of measures were developed. The occupations receiving complete or partial treatment are shown in Figure 8. Those occupations with complete testing are shown on the upper portion of the illustration, while those occupations with partial testing are shown on the lower portion. The primary difference is the development of occupation-specific BARS, and hands-on, procedural knowledge, and rating measures of representative tasks in the 9 occupations shown on the upper half of the illustration. These complement the multiple choice job knowledge measures, "Army-wide" BARS and common task ratings, and leadership, environment, and combat measures developed for all 19 occupations.

## Field Tests

The final step in the development of each criterion measure for the concurrent validation was a field test designed to assess the administrative feasibility, reliability and acceptability of the measures. Field tests were conducted with approximately 150 soldiers in each of the 9 occupations with complete testing programs. These soldiers were about halfway through their first tour. Consequently they were tested at about the same time in their career as the 12,000 soldiers participating in the concurrent validation this summer are being tested. More limited field testing was accomplished with 50 soldiers in each of the 10 additional occupations selected for partial testing. These soldiers were tested at the end of training. This was consistent

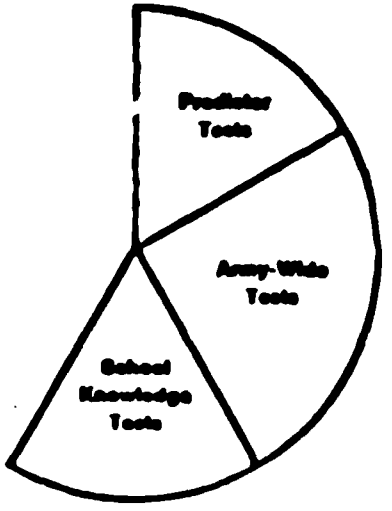
Figure 8

# TESTING COVERAGE




**FULL TREATMENT**
**(BATCH A/BATCH B)**

| MOS | TITLE | MOS | TITLE |
|---|---|---|---|
| 13B | Cannon Crewman | 11B | Infantryman |
| 64C | Motor Transport Oper | 19E | Tank Crewman |
| 71L | Admin Specialist | 31C | Radio TT Oper |
| 95B | Military Police | 63B | Vehicle & Generator Mech |
| | | 91A | Medical Care Specialist |




**PARTIAL TREATMENT**
**(BATCH Z)**

| MOS | TITLE | MOS | TITLE |
|---|---|---|---|
| 12B | Combat Engineer | 54B | Ammunition Spec |
| 16S | MANPADS Crewman | 67N | Utility Helicopter Rpr |
| 27E | Tow/Dragon [illegible] | 76W | Petroleum Supply Spec |
| 51B | Carpentry/[illegible] Spec | 76Y | Unit Supply Spec |
| 54E | Chemical [illegible] Spec | 94B | Food Service Spec |

with the intent that their item-based occupational-specific knowledge test be a measure of training performance. Information obtained from the field tests formed the basis for the selection of the performance measures to be used in the 12,000-soldier concurrent validation of our predictor tests this summer.

The Criterion Measures Task Force

During field testing, the criterion measures just described required about 16 hours for soldiers in the 9 occupations with complete measures. For purposes of the massive concurrent validation, testing time needed to be reduced to 12 hours. For the remaining 10 occupations, measures which required 6 hours, had to be reduced to fit a 4-hour block in the concurrent validation.

A criterion measures task force was established to guide the refinement of the criterion measures based upon the field test data. Their goal was to achieve the testing time reduction while retaining the dimensionality and comprehensiveness of the performance measurement, and the fine-tuning of the tasks, items, and scales suggested by the field test data. The members of the task force are shown in Figure 9. They included members whose objectives were to deal with the overall scientific quality of the measures, as well individuals who were proponents of specific areas of measurement and analysis.

The data analyses from the field tests presented to the task force, as well as the nature of their deliberations, and the guidance they generated, are summarized in the following papers in this symposium. These data have provided intriguing glimpses into the relation-

Figure 9

## CRITERION MEASURES TASK FORCE

Jim Harris - Chair

| | |
|---|---|
| Cross-Task Scientific Affairs | John Campbell<br>Tom Cook<br>Kent Eaton<br>Bob Sadacca<br>Mary Tenopyr |
| Training Measures | Greg Davis |
| Army Wide Measures | Wally Borman |
| Job Specific Measures | Charlotte Campbell<br>Jo Edwards<br>Mike Rumsey |
| Analysis | Karen Mitchell<br>Paul Rossmeissl<br>Laurie Wise |

ships among the different aspects of the performance space, as indexed through different measurement modalities, and of the latent structure of that space.

# PROBLEMS, ISSUES, AND RESULTS IN THE DEVELOPMENT OF TEMPERAMENT, BIOGRAPHICAL, AND INTEREST MEASURES

Leaetta M. Hough, Bruce N. Barge, Janis S. Houston,
Matt K. McGue, and John D. Kamp
Personnel Decisions Research Institute

August 1985

Paper presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

## Author Notes

This paper was prepared as part of a symposium on "Expanding the Measurement of Predictor Space for Military Enlisted Jobs," presented at the annual meeting of the American Psychological Association, August, 1985. Each of the papers discusses a different aspect of developing a set of predictor measures for the Army's Project A, an effort designed to improve the selection, classification and utilization of enlisted personnel. 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 is being funded by the U.S. 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 or policies of the Army Research Institute or the Department of the Army.

## Problems, Issues, and Results in the Development of Temperament, Biographical, and Interest Measures

### Rationale

The Armed Services Vocational Aptitude Battery, ASVAB, has been used with considerable success to predict performance in training; but its usefulness in predicting on-the-job performance is not as well-documented. One objective of the Army Research Institute's Project A, aside from providing data for such documentation, is to expand the type and number of predictors used to select and place Army enlisted personnel.

Temperament, biographical and interest measures were included in the Project A effort because of their potential for predicting different aspects of on-the-job performance and Army-wide criteria such as Effort, Following Regulations and Orders, Leadership, and Self Control.

### Strategy and Results

The approach we used to develop temperament, biographical, and interest measures is outlined in Attachment 1. Our first step was to identify useful constructs, constructs that were likely to predict relevant criteria.

#### Literature Review

We began with a literature review of civilian studies going back to 1960 and of military studies going back to the 1940s. A summary of the criterion-related validity findings, as shown in Tables 1, 2, 3, shows that interest measures have been useful for predicting training criteria, job proficiency criteria, and job involvement criteria with median validities in the high 20s (Table 1). Table 2 shows that biographical inventories have been useful for predicting training, job proficiency, job involvement, and adjustment criteria, with the median validities in the 20s and 30s. Table 3 shows the validities for the temperament constructs. As can be

seen, one or more of the following temperament constructs, Adjustment, Dependability, Achievement, and Locus of Control, appear to predict all but one of the criteria, and that one, Job Involvement, is predicted by interest and biographical inventories.

The present conclusions are similar to those reported by Ghiselli (1973). Ghiselli organized the studies he reviewed according to occupational category and only summarized findings for temperament measures when he judged the temperament construct as pertinent or relevant to a particular occupational category. If we had also categorized our studies according to occupational family such as managerial, clerical, service, trades and craft, sales, protective and industrial as Ghiselli did, perhaps we would have had even more positive findings than are reported here.

Our conclusions do, however, differ from those of the Guion and Gottier (1965) review of criterion-related validity studies. These authors concluded that temperament measures were not very useful for predicting work-relevant criteria. The reason for the difference is that Guion and Gottier did not summarize previous studies according to either a taxonomy of temperament constructs or a taxonomy of criteria.

Table 1

Summary of Criterion-Related Validities of Interest Inventories

| Type of Criterion | Number of Studies | Median r |
|---|---|---|
| Educational | --- | --- |
| Training | 13 | .28 |
| Job Proficiency | 14 | .25 |
| Job Involvement | 21 | .28 |
| Adjustment | --- | --- |

NOTE: Median correlations greater than .20 are indicated by a box.

Table 2

Summary of Criterion-Related Validities of Biographical Inventories

| Type of Criterion | Number of Studies | Median r |
|---|---|---|
| Educational | --- | --- |
| Training | 18 | .22 |
| Job Proficiency | 48 | .38 |
| Job Involvement | 33 | .31 |
| Adjustment | 6 | .23 |

NOTE: Median correlations greater than .20 are indicated by a box.

Table 3

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

| Temperament Construct | Type of Criterion: Educational | Training | Job Proficiency | Job Involvement | Adjustment |
|---|---|---|---|---|---|
| Potency (Surgency) | .06 (42)[b] | .13 (36) | .07 (65) | .04 (13) | -.17 (31) |
| Adjustment | .14 (43) | .19 (28) | .11 (65) | .17 (16) | **-.33** (52) |
| Agreeableness (Likeability) | .03 ( 9) | .08 ( 5) | .03 (22) | -.02 ( 5) | -.03 ( 5) |
| Dependability | .13 (24) | .12 (20) | .11 (49) | .14 (15) | **-.43** (40) |
| Intellectance (Culture) | .17 ( 6) | .19 ( 5) | .01 (16) | -.09 ( 9) | .18 ( 3) |
| Affiliation | -.03 ( 5) | --- --- | -.02 ( 6) | .09 ( 4) | -.07 ( 4) |
| Achievement | **.30** ( 8) | **.33** ( 4) | **.24** ( 4) | --- --- | **-.33** ( 5) |
| Masculinity | -.16 ( 8) | .09 ( 3) | .10 (10) | .03 ( 4) | -.13 (11) |
| Locus of Control | **.32** ( 1) | **.29** ( 2) | **.25** ( 7) | --- --- | --- --- |
| Unclassified Military Scales | --- --- | .18 ( 8) | .18 (25) | --- --- | **-.22** (20) |

[a] Medians are reported as the summary index.

[b] The number in parenthesis is the number of correlations on which the median is based.

NOTE: Median correlations greater than .20 are indicated by a box.

Expected True Validities

Our next step was to obtain expert judgments about expected true criterion-related validities of our predictor constructs for Army criteria. As can be seen in Table 4, all of the temperament scales developed for this project have good potential for predicting Army-wide performance criteria.

Selection of Constructs

The constructs that we determined were "good bets" appear in List 1. They are the constructs we intended to measure in our initial experimental battery of new tests, called the Pilot Trial Battery.

Development of Measures

The first issue we had to resolve was the homogeneity/heterogeneity issue. We decided to specify components of the constructs and to write reasonably homogeneous items to measure the components. We called these components scales.

The second issue we had to resolve was the item and response format. We had previously factor analyzed approximately 1800 soldiers' responses to off-the-shelf temperament, biographical, and interest inventories and found that the biographical items split and formed factors with either temperament or interest items, but that temperament and interest items did not jointly load on any factors (Hough, 1984). We concluded that biodata type items did not tap unique constructs. In Wernimont and Campbell's (1968) terminology of signs versus samples, biodata items are samples and self-perception items are samples. We, therefore, decided to develop two inventories and to use both biodata and self-perception items in each. The two inventories are the ABLE (Assessment of Background and Life Experiences) and the AVOICE (Army Vocational Interest Career Examination). The scales that we included in the ABLE appear in List 2, the scales for the AVOICE in List 3. We developed four response validity scales that

Table 4

ESTIMATED VALIDITIES FOR ABLE* SCALES BASED ON EXPERT JUDGMENT

| Army-Wide Performance Criteria: | Emotional Stability | Self-Esteem | Cooperativeness | Conscientiousness | Nondelinquency | Traditional Values | Work Orientation | Internal Control | Energy Level | Dominance | Physical Condition |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Technical Knowledge/Skill | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| Initiative/Effort | 24 | 23 | 24 | 32 | 27 | 22 | 46 | 32 | 38 | 21 | 17 |
| Following Regulations and Orders | 25 | 16 | 40 | 38 | 45 | 40 | 32 | 17 | 16 | 10 | 12 |
| Integrity | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| Leading and Supporting | 31 | 26 | 41 | 26 | 31 | 20 | 31 | 24 | 27 | 28 | 15 |
| Maintaining Assigned Equipment | 22 | 16 | 17 | 46 | 27 | 19 | 41 | 24 | 19 | 10 | 08 |
| Maintaining Living/Work Areas | | | | | | | | | | | |
| Military Appearance | 20 | 17 | 28 | 32 | 42 | 38 | 30 | 17 | 14 | 16 | 15 |
| Physical Fitness | 17 | 20 | 10 | 21 | 16 | 14 | 21 | 23 | 35 | 18 | 55 |
| Self-Development | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| Self-Control | 32 | 20 | 18 | 43 | 44 | 34 | 38 | 28 | 16 | 12 | 14 |

NOTE: The decimals have not been included; they were two digits to the left.

*ABLE = Assessment of Background and Life Experiences; the inventory that contains the temperament/biodata scales.

List 1

Constructs in the Pilot Trial Battery

- Potency (Surgency)
- Adjustment
- Agreeableness (Likeability)
- Dependability
- Achievement
- Locus of Control
- Physical Condition
- Social Interests
- Realistic Interests
- Investigative Interests
- Enterprising Interests
- Artistic Interests
- Conventional Interests
- Expressed Interests
- Organizational Climate/Environment Preferences

List 2

ABLE Scales Organized by Construct

CONTENT SCALES:

Potency (Surgency)

- Dominance
- Energy Level

Adjustment

- Emotional Stability
- Self Esteem

Agreeableness (Likeability)

- Cooperativeness

Dependability

- Nondelinquency
- Traditional Values
- Conscientiousness

Achievement

- Work Orientation

Locus of Control

- Internal Control

Physical Condition

- Physical Condition

RESPONSE VALIDITY SCALES:

- Non-Random Responses
- Unlikely Virtues (Social Desirability)
- Poor Impression
- Self Knowledge

List 3

AVOICE Scales Organized by Construct

Realistic Interests

. Basic Interest Item

* . Mechanics

* . Heavy Construction

* . Electronics

* . Electronic Communication

* . Drafting

* . Law Enforcement

* . Audiographics

. Agriculture

* . Outdoors

* . Marksman

* . Infantry

* . Armor/Cannon

* . Vehicle Operator

* . Adventure

Conventional Interests

. Basic Interest Item

* . Office Administration

* . Supply Administration

* . Food Service

Social Interests

. Basic Interest Item

* . Teaching/Counseling

List 3 (Continued)

Investigative Interests

. Basic Interest Item

* . Medical Services

* . Mathematics

* . Science/Chemical

* . Automated Data Processing

Enterprising Interests

. Basic Interest Item

* . Leadership

Artistic Interests

. Basic Interest Item

* . Aesthetics

Organizational Climate/Environment Preferences

* . Achievement Preferences

* . Safety Preferences

* . Comfort Preferences

* . Status Preferences

* . Altruism Preferences

* . Autonomy Preferences

Expressed Interests

. Expressed Interests

NOTE: All the above scales were included in the pilot trial battery.

* Indicates scales included in trial battery.

were included in the ABLE: Non-Random Responses, Unlikely Virtues (Social Desirability), Poor Impression, and Self-Knowledge.

Evaluation of Scales

Sensitivity of Item Content. An important concern for the Army is the sensitivity, or lack of sensitivity, of the temperament items. Psychologists and military personnel both reviewed the items to ensure that the content of the items was acceptable. Several revisions were made in response to their suggestions.

Psychometric Characteristics. The temperament and interest measures of the pilot trial battery were administered to soldiers at Ft. Campbell in May 1984, at Ft. Lewis in June 1984, and Ft. Knox in September 1984. Each time the means, standard deviations, item response distributions, scale score distributions, item-total scale correlations, and internal consistency indices (alpha coefficients) were examined and the information used to revise the items and scales. At Fort Knox, we also retested about 130 soldiers. Much of Ft. Knox summary data are presented in Tables 5, 6, and 7. As can be seen, the median alpha coefficient (internal consistency) for the ABLE content scales is .84, with a range of .70-.87; the median test-retest correlation for the ABLE content scales is .79, with a range of .68-.83. At retest or second testing, the soldiers apparently responded in a more random way. The response validity scale, Non-Random Responses, detected it and, consequently, the correlation between first and second testing was low, .37. The median alpha coefficient (internal consistency) for the AVOICE scales is .86 with a range of .68 to .96. The median test-retest correlation for the AVOICE scales is .76, with a range of .56 to .86. Clearly, the scales are internally consistent and yield stable estimates of a person's score. These data also suggested that the scales could be reduced in length without sacrificing much precision in measurement.

Table 5

Pilot Trial Battery (PTB)

Summary of ABLE Scale Score Characteristics
(N=276 except where otherwise noted)

Ft. Knox

| Scale | # Items | Mean | SD | Alpha | Test-Retest [3] r | Median Item-Scale r |
|---|---|---|---|---|---|---|
| **Content Scales** | | | | | | |
| 1. Emotional Stability | 29 | 64.94 | 8.27 | .86 | .68[1] | .44 |
| 2. Self-Esteem | 15 | 35.10 | 5.25 | .83 | .81 | .54 |
| 3. Cooperativeness | 24 | 54.08 | 6.09 | .77 | .69 | .42 |
| 4. Conscientiousness | 21 | 48.92 | 5.90 | .81 | .73 | .43 |
| 5. Nondelinquency | 24 | 55.44 | 7.23 | .84 | .81 | .46 |
| 6. Traditional Values | 16 | 37.23 | 4.60 | .70 | .74 | .45 |
| 7. Work Orientation | 27 | 61.20 | 7.93 | .85 | .80 | .47 |
| 8. Internal Control | 21 | 50.31 | 6.14 | .79 | .75 | .43 |
| 9. Energy Level | 25 | 57.14 | 7.11 | .85 | .79 | .47 |
| 10. Dominance | 16 | 35.46 | 6.13 | .86 | .83 | .56 |
| 11. Physical Condition | 9 | 31.06 | 7.53 | .87 | .81 | .72 |
| **Response Validity Scales** | | | | | | |
| 12. Unlikely Virtues | 12 | 16.60 | 3.39 | .68 | .62 | .53 |
| 13. Self-Knowledge | 13 | 29.64 | 3.54 | .62 | .71 | .41 |
| 14. Non-Random Response[2] | 8 | 7.68 | .71 | .56 | .37 | .45 |
| 16. Poor Impression | 24 | 1.54 | 1.86 | .61 | .56 | .33 |

[1] N=109 for Test-Retest Correlations.

[2] N=281. Statistics reported for this scale are based on sample edited for Overall M.D. only. "Passing" score on Non-Random Response Scale $\geq$ 6.

[3] Test-Retest internal was two weeks.

Table 6

PTB: ABLE Test-Retest[1]

Ft. Knox

| | Mean Time 1 (N=276) | Mean Time 2 (N=109) | Effect Size $\frac{X_2 - X_1}{SD_1}$ |
|---|---|---|---|
| Content Scales: | | | |
| 1. Emotional Stability | 64.9 | 65.1 | .02 |
| 2. Self-Esteem | 35.1 | 34.8 | -.05 |
| 3. Cooperativeness | 54.1 | 54.3 | .04 |
| 4. Conscientiousness | 48.9 | 48.3 | -.10 |
| 5. Nondelinquency | 55.4 | 55.6 | .02 |
| 6. Traditional Values | 37.2 | 37.9 | .15 |
| 7. Work Orientation | 61.2 | 60.7 | -.07 |
| 8. Internal Control | 50.3 | 50.2 | -.01 |
| 9. Energy Level | 57.1 | 57.0 | -.01 |
| 10. Dominance | 35.5 | 34.9 | -.09 |
| 11. Physical Condition | 31.1 | 30.4 | -.09 |
| Response Validity Scales: | | | |
| 12. Unlikely Virtues | 16.6 | 17.5 | .27 |
| 13. Self-Knowledge | 29.6 | 29.0 | -.18 |
| 14. Non-Random Response[2] | 7.7 | 7.2 | -.65 |
| 16. Poor Impression | 1.5 | 1.2 | -.18 |

[1]Test-Retest interval was two weeks.

[2]Based on sample edited for M.D. only; $N_1$=281 and $N_2$=121.

Table 7

Pilot Trial Battery (PTB)

Summary of AVOICE Scale Score Characteristics
(N=270 except where otherwise noted)

Ft. Knox

| Scale | # Items | Mean | SD | Alpha | Test-Retest r | Median Item-Scale r |
|---|---|---|---|---|---|---|
| 1. Marksman | 5 | 15.80 | 4.37 | .79 | .77[1] | .75 |
| 2. Agriculture | 5 | 14.07 | 3.99 | .68 | .69 | .70 |
| 3. Mathematics | 5 | 15.11 | 4.37 | .82 | .76 | .79 |
| 4. Aesthetics | 5 | 14.26 | 4.17 | .77 | .72 | .74 |
| 5. Leadership | 6 | 20.33 | 4.70 | .81 | .56 | .74 |
| 6. Electronic Communication | 7 | 21.13 | 5.73 | .92 | .78 | .72 |
| 7. Automated Data Processing | 7 | 23.35 | 6.56 | .88 | .81 | .81 |
| 8. Teacher/Counseling | 7 | 22.85 | 5.53 | .82 | .73 | .73 |
| 9. Drafting | 7 | 21.47 | 6.12 | .85 | .74 | .77 |
| 10. Audiographics | 7 | 23.80 | 5.68 | .82 | .76 | .70 |
| 11. Armor/Cannon | 8 | 22.43 | 6.57 | .83 | .74 | .69 |
| 12. Vehicle/Equipment Operator | 10 | 28.07 | 7.79 | .86 | .69 | .70 |
| 13. Outdoors | 9 | 31.71 | 6.41 | .79 | .69 | .66 |
| 14. Infantry | 10 | 29.12 | 7.13 | .81 | .78 | .65 |
| 15. Science/Chemical Operations | 11 | 29.35 | 8.93 | .89 | .79 | .71 |
| 16. Supply Administration | 13 | 34.99 | 10.44 | .92 | .82 | .75 |
| 17. Office Administration | 16 | 45.18 | 13.20 | .94 | .86 | .73 |
| 18. Law Enforcement | 16 | 48.12 | 11.84 | .88 | .78 | .63 |
| 19. Mechanics | 16 | 50.01 | 14.68 | .95 | .80 | .80 |
| 20. Electronics | 20 | 59.96 | 17.48 | .96 | .74 | .77 |
| 21. Heavy Construction/Combat | 23 | 65.75 | 17.90 | .94 | .76 | .70 |
| 22. Medical Services | 24 | 68.46 | 18.79 | .95 | .84 | .69 |
| 23. Food Service | 17 | 42.24 | 11.16 | .89 | .71 | .64 |

[1]N=127 for Test-Retest Correlations.

Structural Characteristics. At Ft. Campbell we also administered four "off-the-shelf" scales that measured four of the same temperament constructs that we were trying to measure in the pilot trial battery. (These "off-the-shelf" scales had been included in an earlier battery, called the Preliminary Battery, which is discussed in Hough et al., 1984.) Table 8 shows the correlations of the ABLE scales and constructs with these four Preliminary Battery scales. As can be seen, the ABLE scales do correlate most highly with the target Preliminary Battery scale, demonstrating good convergent and discriminant validity.

We separately factor analyzed the ABLE content scales and AVOICE scales (using the Ft. Knox sample data) and found in both cases the two factor solution best summarized the data. As shown in Table 9 and 10, the temperament factors were labeled Personal Impact and Dependability; the two interest factors were labeled Combat-Related and Combat-Support.

Unique contribution of ABLE and AVOICE to predictor battery. We examined the scales for their potential for providing incremental validity to the predictor battery. Each scale was compared in terms of its uniqueness in relation to the ASVAB and the psychomotor measures and cognitive measures in the pilot trial battery. As shown in Tables 11, 12, and 13, the ABLE content scales and the AVOICE share very little of the same variance as the ASVAB or the psychomotor and cognitive measures included in the pilot trial battery.

Response Sets. In addition to conducting analyses designed to detect response sets such as true vs false, extreme responding, and inattention to the direction, positive or negative, of the item stem, we also developed four response validity scales: Non-Random Responses, Unlikely Virtues (Social Desirability), Poor Impression, and Self Knowledge. We also conducted a study, including an experiment, on intentional distortion

Table 8

Fort Campbell Pilot Test, May 1984
Correlations Between ABLE Constructs and Scales
(Pilot Trial Battery) and POI
(Personal Opinion Inventory - Preliminary Battery)
Marker Variables

| | POI Scale | | | |
|---|---|---|---|---|
| ABLE Construct | DPQ Stress Reaction | DPQ Social Potency | Rotter Locus of Control | CPI Socialization |
| Adjustment | [-.63] | .38 | .37 | .39 |
| Leadership | -.33 | [.64] | .39 | .31 |
| Locus of Control | -.32 | .26 | [.67] | .60 |
| Dependability | -.34 | .20 | .43 | [.62] |

| | POI Scale | | | |
|---|---|---|---|---|
| ABLE Scale | DPQ Stress Reaction | DPQ Social Potency | Rotter Locus of Control | CPI Socialization |
| Emotional Stability | [-.70] | .32 | .30 | .32 |
| Dominance | -.24 | [.67] | .18 | .22 |
| Locus of Control | -.32 | .26 | [.67] | .60 |
| Nondelinquency | -.34 | .16 | .32 | [.62] |

Note: "Marker" correlations are indicated by a box.

N = 38

Table 9

Pilot Trial Battery: ABLE

Factor Analysis[1]

Ft. Knox

| | I Impact (Vitality/Energy) | II Socialization (Dependability/Discipline) | $h^2$ |
|---|---|---|---|
| Self Esteem | .80 | .30 | .73 |
| Energy Level | .73 | .46 | .74 |
| Dominance (Leadership) | .72 | .13 | .54 |
| Emotional Stability | .67 | .26 | .52 |
| Work Orientation | .67 | .51 | .71 |
| Nondelinquency | .20 | .81 | .70 |
| Traditional Values | .19 | .73 | .57 |
| Conscientiousness | .39 | .72 | .67 |
| Cooperativeness | .46 | .60 | .57 |
| Internal Control | .44 | .50 | .44 |
| | | | 6.19 |

[1] Principal factor analysis, varimax rotation

N = 276

Table 10

Pilot Trial Battery: AVOICE
Factor Analysis[1]
Ft. Knox

| Scale | I Combat Support* | II Combat Related** | $h^2$ |
|---|---|---|---|
| Office Administration | .85 | -.13 | .73 |
| Supply Administration | .78 | .11 | .62 |
| Teacher/Counseling | .76 | .11 | .59 |
| Mathematics | .74 | .09 | .55 |
| Medical Services | .73 | .18 | .57 |
| Automated Data Processing | .71 | .10 | .51 |
| Audiographics | .64 | .17 | .44 |
| Electronic Communication | .64 | .36 | .54 |
| Science/Chemical Operations | .61 | .43 | .55 |
| Aesthetics | .61 | .04 | .37 |
| Leadership | .58 | .35 | .46 |
| Food Service | .54 | .19 | .33 |
| Drafting | .54 | .34 | .41 |
| Infantry | .10 | .85 | .74 |
| Armor/Cannon | .13 | .84 | .73 |
| Heavy Construction/Combat | .17 | .84 | .73 |
| Outdoors | .02 | .74 | .55 |
| Mechanics | .17 | .74 | .58 |
| Marksman | .05 | .73 | .54 |
| Vehicle/Equipment Operator | .17 | .73 | .56 |
| Agriculture | .18 | .64 | .44 |
| Law Enforcement | .27 | .61 | .44 |
| Electronics | .45 | .57 | .52 |
| | | | 12.49 |

[1]Principal factor analysis, varimax rotation
*Conventional, Social, Investigative, Enterprising, Artistic
**Realistic

N = 270

Table 11

Pilot Trial Battery (PTB)
Uniqueness Analyses of Content ABLE Scales
Ft. Knox

| Scale | # Items | Alpha (N=276) | Test-Retest r (N=109) | ABLE Adj $R^2$ (N=207) | ASVAB Adj $R^2$ (N=183) | ASVAB $U^2$ Using Alpha (N=183) | ASVAB $U^2$ Using T-R (N=183) |
|---|---|---|---|---|---|---|---|
| Emotional Stability | 29 | .86 | .68 | .52 | .05 | .81 | .63 |
| Self-Esteem | 15 | .83 | .81 | .70 | .03 | .80 | .78 |
| Cooperativeness | 24 | .77 | .69 | .54 | .00 | .77 | .69 |
| Conscientiousness | 21 | .81 | .73 | .64 | .03 | .78 | .70 |
| Nondelinquency | 24 | .84 | .81 | .63 | .02 | .82 | .79 |
| Traditional Values | 16 | .70 | .74 | .50 | .01 | .69 | .73 |
| Work Orientation | 27 | .85 | .80 | .71 | .03 | .82 | .77 |
| Internal Control | 21 | .79 | .75 | .48 | .04 | .75 | .71 |
| Energy Level | 25 | .85 | .79 | .72 | .05 | .80 | .74 |
| Dominance | 16 | .86 | .83 | .50 | .00 | .86 | .83 |
| Physical Condition | 9 | .87 | .81 | .11 | .00 | .87 | .81 |

Table 12

Pilot Trial Battery (PTB)
Uniqueness Analyses of AVOICE Occupational Scales
Ft. Knox

| Scale | # Items | Alpha (N=270) | Test-Retest r (N=127) | ASVAB Adj $R^2$ (N=149) | ASVAB $U^2$ Using Alpha (N=149) | ASVAB $U^2$ Using T-R (N=149) |
|---|---|---|---|---|---|---|
| Marksman | 5 | .79 | .77 | .20 | .59 | .57 |
| Agriculture | 5 | .68 | .69 | .06 | .62 | .63 |
| Mathematics | 5 | .82 | .76 | .02 | .80 | .74 |
| Aesthetics | 5 | .77 | .72 | .08 | .69 | .64 |
| Leadership | 6 | .81 | .56 | .00 | .81 | .56 |
| Electronic Communication | 7 | .92 | .78 | .01 | .91 | .77 |
| Automated Data Processing | 7 | .88 | .81 | .00 | .88 | .81 |
| Teacher/Counseling | 7 | .82 | .73 | .00 | .82 | .73 |
| Drafting | 7 | .85 | .74 | .07 | .78 | .67 |
| Audiographics | 7 | .82 | .76 | .00 | .82 | .76 |
| Armor/Cannon | 8 | .83 | .74 | .11 | .72 | .63 |
| Vehicle/Equipment Operator | 10 | .86 | .69 | .14 | .72 | .55 |
| Outdoors | 9 | .79 | .69 | .16 | .63 | .53 |
| Infantry | 10 | .81 | .78 | .13 | .68 | .65 |
| Science/Chemical Operations | 11 | .89 | .79 | .01 | .88 | .78 |
| Supply Administration | 13 | .92 | .82 | .00 | .92 | .82 |
| Office Administration | 16 | .94 | .86 | .03 | .91 | .83 |
| Law Enforcement | 16 | .88 | .78 | .02 | .86 | .76 |
| Mechanics | 16 | .95 | .80 | .32 | .63 | .48 |
| Electronics | 20 | .96 | .74 | .14 | .82 | .60 |
| Heavy Construction/Combat | 23 | .94 | .76 | .21 | .73 | .55 |
| Medical Services | 24 | .95 | .84 | .00 | .95 | .84 |
| Food Service | 17 | .89 | .71 | .02 | .87 | .69 |
| Adventure | 14 | .96 | .86 | .26 | .70 | .60 |

Table 13

Summary of Overlap with Other PTB Measures

Ft. Knox

1. Between ABLE and PTB Cognitive Tests:

   . Only 19%, 29 of 150 correlations, are significant at $p \leq .05$.

   . The highest correlation is .23.

2. Between ABLE and PTB Computer Measures:

   . Only 17%, 48 of 285 correlations, are significant at $p \leq .05$.

   . The highest correlation is .24.

3. Between AVOICE and PTB Cognitive Tests:

   . Only 36%, 128 of 360 correlations, are significant at $p \leq .05$.

   . The highest correlation is .32.

4. Between AVOICE and PTB Computer Measures:

   . Only 15%, 105 of 684 correlations, are significant at $p \leq .05$.

   . The highest correlation is .30.

(faking) of responses. We gathered data 1) from soldiers instructed, at different times, to distort their responses and to be honest (experimental data gathered at Ft. Bragg); 2) from soldiers who were simply responding to our inventories with no particular directions (data gathered at Ft. Knox, in another type of "honest" condition); and 3) from candidates at the Military Entrance Processing Station (MEPS) who wanted to be accepted into the Army.

The purpose of the faking study was to:

1. Determine extent to which soldiers can distort their responses to temperament and interest inventories when instructed to do so. (Compare data from Ft. Bragg faking conditions with Ft. Bragg and Ft. Knox honest conditions.)
2. Determine extent to which the ABLE response validity scales detect such intentional distortion. (Compare response validity scales in Ft. Bragg honest and faking conditions.)
3. Determine extent to which ABLE validities scales can be used to correct or adjust scores on substantive scales.
4. Determine extent to which distortion is a problem in an applicant setting. (Compare MEPS data with Ft. Bragg and Ft. Knox data.)

The participants in the experimental group were 245 enlisted soldiers in the 82nd airborne brigade at Fort Bragg. Comparison samples were MEPS (Army) candidates (N=125) and Ft. Knox soldiers earlier described (N=230).

Procedure and Design

Four faking conditions were created:

. Fake Good on the ABLE

. Fake Bad on the ABLE

. Fake combat on the AVOICE

. Fake non-combat on the AVOICE

Two honest conditions were created:

. Honest on the ABLE

. Honest on the AVOICE

The significant part of the instructions for the six conditions were as follows:

ABLE - Fake Good

Imagine you are at the Military Entrance Processing Station (MEPS) and you want to join the Army. Describe yourself in a way that you think will ensure that the Army selects you.

ABLE - Fake Bad

Imagine you are at the Military Entrance Processing Station (MEPS) and you do not want to join the Army. Describe yourself in a way that you think will ensure that the Army does not select you.

ABLE - Honest

You are to describe yourself as you really are.

AVOICE - Fake Combat

Imagine you are at the Military Entrance Processing Station (MEPS). Please describe yourself in a way that you think will ensure that you are placed in an occupation in which you are likely to be exposed to combat during a wartime situation.

AVOICE - Fake Non-combat

Imagine you are at the Military Entrance Processing Station (MEPS). Please describe yourself in a way you think will ensure that you are placed in an occupation in which you are unlikely to be exposed to combat during a wartime situation.

AVOICE - Honest

You are to describe yourself as you really are.

The design was repeated measures with faking and honest conditions counter-balanced. Thus, approximately half the experimental group, 124 soldiers, completed the inventories honestly in the morning and faked in the afternoon, while the other half (121) completed the inventories honestly in the afternoon and faked in the morning.

The design and numbers of soldiers from whom we gathered the intentional faking data appear in Table 14.

In summary, then, a 2 x 2 x 2 fixed-factor completely crossed experimental design was used. The within-subjects factor consisted of two levels (honest responses and faked responses). The first between subjects factor consisted of the following two levels: fake good/want combat and fake bad/do not want combat. Order was manipulated in the second between subjects factor such that the following two levels were produced: faked responses then honest responses and honest responses then faked responses.

Summary of Faking Study Results - Temperament Inventory

We performed a multivariate analysis of variance (MANOVA) on the experimental data (Ft. Bragg data); Table 15 shows the findings for the interactions, the sources of variance most relevant to the question of whether soldiers can or cannot intentionally distort their responses. As can be seen, all the fake x set interactions are significant, indicating that soldiers can, when instructed to do so, distort their responses.

Table 15 also shows that the overall test of significance for the fake x set x order interaction effect is statistically significant for the response validity scales and marginally significant for the content scales. These results indicate that order of conditions in which the participant

Table 14

Faking Experiment
ABLE and AVOICE
Fort Bragg, September 1984

AVOICE/ABLE Counts

Monday

AM: Honest AVOICE
Honest ABLE
N=64

PM: Fake Combat AVOICE
Fake Good ABLE
N=62

(62) Complete Sets

Tuesday

AM: Honest AVOICE
Honest ABLE
N=62

PM: Fake Noncombat AVOICE
Fake Bad ABLE
N=62

(62) Complete Sets

Wednesday

AM: Fake Combat AVOICE
Fake Good ABLE
N=63

PM: Honest AVOICE
Honest ABLE
N=61

(61) Complete Sets

Thursday

AM: Fake Noncombat AVOICE
Fake BAD ABLE
N=61

PM: Honest AVOICE
Honest ABLE
N=60

(60) Complete Sets

Table 15

Ft. Bragg Fakability Study
MANOVA Results
ABLE Scales

| Type and Name of Scale | Interactions | |
|---|---|---|
| Response Validity Scales | Fake x Set | Fake x Set x Order |
| Overall | S | S |
| Social Desirability (Unlikely Virtues) | S | S |
| Self-Knowledge | S | NS |
| Non-Random Response | S | NS |
| Poor Impression | S | NS |
| Content Scales | | |
| Overall | S | NS* |
| Emotional Stability | S | --- |
| Self-Esteem | S | --- |
| Cooperativeness | S | --- |
| Conscientiousness | S | --- |
| Non-Delinquency | S | --- |
| Traditional Values | S | --- |
| Work Orientation | S | --- |
| Internal Control | S | --- |
| Energy | S | --- |
| Dominance (Leadership) | S | --- |

Note: S = significant, $p \leq .01$
NS = nonsignificant, $p > .01$
* = marginally significant, $.05 \leq p > .01$
Sample Size for Response Validity Scales is 219
Sample Size for Content Scales is 208

completed the ABLE affected the results. Table 16 shows in greater detail the effects of intentional distortion; it shows the mean scores for the various experimental conditions. This tabnle and the remaining tables showing Ft. Bragg ABLE results report the values for the first administration of the particular condition.

Another research question was the extent to which our response validity scales detected intentional distortion. As can be seen in Table 17, the response validity scale Social Desirability (Unlikely Virtues) detects faking good on the ABLE; the response validity scales Non-Random Response, Poor Impression and Self-Knowledge detect faking bad. According to these data, the soldiers responded more randomly, created a poorer impression, and reported that they knew themselves less well when told to describe themselves in a way that would increase the likelihood that they would not be accepted into the Army.

We also examined the extent to which we could use the response validity scales Social Desirability and Poor Impression to adjust ABLE content scale scores for faking good and faking bad. We regressed out Social Desirability from the content scales in the fake good condition and Poor Impression from the content scales in the fake bad condition.

Table 18 shows the mean differences in content scales before and after regressing out Social Desirability and Poor Impression. Clearly, these response validity scales can be used to adjust content scales; however, two important unknowns remain: Do the adjustment formula developed on these data cross validate and do they increase criterion-related validity?

Another very important finding of the present study is that applicants do not tend to distort their responses to the ABLE. Table 19 shows mean scores for applicants (MEPS) and the two honest conditions, Ft. Bragg and Ft. Knox. Even though the applicants are trying not to create a poor

Table 16
Able Content Scales
Examination of Honesty and Faking Effects

| Scale | Honest* | | | Fake Good* | | | Fake Bad* | | | Estimated Effect Size Honest vs Good | Estimated Effect Size Honest vs Bad |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | M | S.D. | N | M | S.D. | N | M | S.D. | | |
| Emotional Stability | 103 | 66.2 | 7.8 | 54 | 70.3 | 10.2 | 54 | 50.1 | 10.8 | -.46 | 1.73 |
| Self-Esteem | 103 | 34.8 | 4.7 | 54 | 38.2 | 5.4 | 54 | 22.2 | 5.8 | -.67 | 2.40 |
| Cooperativeness | 103 | 53.4 | 6.2 | 54 | 55.5 | 8.8 | 54 | 36.7 | 10.4 | -.28 | 2.01 |
| Conscientiousness | 103 | 46.4 | 5.7 | 54 | 49.6 | 8.4 | 54 | 31.7 | 8.7 | -.45 | 2.04 |
| Non-delinquency | 103 | 53.3 | 6.1 | 54 | 54.8 | 10.2 | 54 | 36.8 | 9.6 | -.18 | 2.10 |
| Traditional Values | 103 | 36.8 | 4.6 | 54 | 38.7 | 6.5 | 54 | 23.6 | 6.1 | -.34 | 2.47 |
| Work Orientation | 103 | 59.6 | 7.6 | 54 | 64.7 | 10.3 | 54 | 40.8 | 11.7 | -.57 | 1.02 |
| Internal Control | 103 | 49.6 | 6.4 | 54 | 50.9 | 8.2 | 54 | 35.6 | 8.9 | -.18 | 1.83 |
| Energy Level | 103 | 57.6 | 6.7 | 54 | 61.4 | 9.1 | 54 | 37.9 | 1.4 | -.48 | 4.86 |
| Dominance (Leadership) | 103 | 35.5 | 5.8 | 54 | 40.3 | 5.6 | 54 | 24.5 | 6.6 | -.84 | 1.77 |

*Mean scores are based on persons who responded to this condition first.

Table 17

ABLE Response Validity Scales:
Effects of Honest* and Faking* Conditions

| 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 | 15.7 | 3.1 | 57 | 20.1 | 5.8 | 56 | 17.8 | 4.8 | -1.0 | - .5 |
| Self-Knowledge | 109 | 29.6 | 3.6 | 57 | 29.7 | 4.1 | 56 | 21.8 | 5.2 | - .00 | 1.8 |
| Non-Random Response | 109 | 7.6 | .9 | 57 | 7.0 | 1.8 | 56 | 2.8 | 2.2 | .4 | 3.1 |
| Poor Impression | 109 | 1.5 | 2.1 | 57 | 1.7 | 2.2 | 56 | 14.6 | 7.9 | - .1 | -2.6 |

*Values are based on the sample that completed the questionnaires under the condition of interest first.

Table 18

Effects of Regressing Out Response
Validity Scales (Social Desirability and Poor Impression)
on Faking Condition ABLE Content Scale Scores

| | Fake Good Condition | Fake Bad Condition |
|---|---|---|
| Content Scales: | | |
| Emotional Stability | | |
| Self Esteem | | |
| Cooperativeness | | |
| Conscientiousness | | |
| Non-delinquency | | |
| Traditional Values | | |
| Work Orientation | | |
| Internal Control | | |
| Energy | | |
| Dominance (Leadership) | | |

Note: Mean differences are between group comparisons for first testing only. For the Fake Good condition they are [mean (Fake) - mean (Honest)]/SD. For the Fake Bad condition they are [mean (Honest) - mean (Fake)]/SD.

Table 19

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. | Degrees of Freedom | F | P |
|---|---|---|---|---|---|---|---|---|---|---|
| **Response Validity Scales** | | | | | | | | | | |
| Social Desirability (Unlikely Virtues) | 116 | 15.91 | 121 | 16.63 | 276 | 16.60 | 3.21 | 2,510 | 2.15 | .12 |
| Self-Knowledge | 116 | 29.54 | 121 | 28.03 | 276 | 29.64 | 3.63 | 2,510 | 9.10 | .00 |
| Non-Random Response | 116 | 7.58 | 121 | 7.79 | 276 | 7.75 | .64 | 2,510 | 3.73 | .02 |
| Poor Impression | 116 | 1.50 | 121 | 1.05 | 276 | 1.54 | 1.84 | 2,510 | 3.15 | .04 |
| **Content Scales** | | | | | | | | | | |
| Emotional Stability | 112 | 66.22 | 118 | 66.03 | 272 | 65.05 | 7.86 | 2,499 | 1.18 | .31 |
| Self-Esteem | 112 | 34.77 | 118 | 34.04 | 272 | 35.12 | 5.00 | 2,499 | 1.93 | .15 |
| Cooperativeness | 112 | 53.33 | 118 | 54.60 | 272 | 54.19 | 6.05 | 2,499 | 1.34 | .26 |
| Conscientiousness | 112 | 46.37 | 118 | 46.49 | 272 | 48.97 | 5.86 | 2,499 | 12.24 | .00 |
| Non-Delinquency | 112 | 53.24 | 118 | 54.36 | 272 | 55.49 | 6.91 | 2,499 | 4.48 | .01 |
| Traditional Values | 112 | 36.67 | 118 | 36.97 | 272 | 37.28 | 4.50 | 2,499 | .77 | .46 |
| Work Orientation | 112 | 59.71 | 118 | 58.37 | 272 | 61.40 | 7.73 | 2,499 | 6.90 | .00 |
| Internal Control | 112 | 49.48 | 118 | 51.90 | 272 | 50.37 | 6.13 | 2,499 | 4.75 | .01 |
| Energy Level | 112 | 57.56 | 118 | 56.67 | 272 | 57.19 | 6.95 | 2,499 | .48 | .62 |
| Dominance (Leadership) | 112 | 35.54 | 118 | 32.84 | 272 | 35.41 | 6.05 | 2,499 | 8.69 | .00 |
| Physical Condition | 112 | 32.96 | 118 | 28.27 | 272 | 31.08 | 7.49 | 2,499 | 12.10 | .00 |

*Scores are based on persons who responded to the honest condition first.

impression (MEPS mean is 1.05, which is lower than both the Ft. Knox and Ft. Bragg means, 1.54 and 1.50 respectively), they do not score significantly higher on the response validity scale Social Desirability (Unlikely Virtues). Indeed their mean score is lowest on six of the eleven content scales, scales on which it would be desirable to score high rather than low. They score highest on only two content scales and only one is significant, Internal Control. In sum, intentional distortion does not appear to be a problem in an applicant setting. Faking or distortion in a draft situation cannot be assessed in the present U.S. situation.

Overall, the ABLE data from the faking study show that:

1. Soldiers can distort their responses when instructed to do so;
2. The response validity scales detect intentional faking; Social Desirability (Unlikely Virtues) detects faking good and Non-Random Response, Poor Impression, and Self Knowledge detect faking bad;
3. An individual's Social Desirability scale score can be used to adjust his or her content scale scores to reduce variance associated with faking good; an individual's Poor Impression scale score can be used to adjust his or her content scale scores to reduce variance associated with faking bad; and
4. Faking or distortion does not appear to be a problem in an applicant setting.

Summary of Faking Study Results - Interest Inventory

Soldiers can distort their responses when instructed to do so. We divided the interest scales into the two groups, combat-related and combat-support, that emerged when we factor analyzed the AVOICE Ft. Knox data. Again, we performed a multivariate analysis of variance (MANOVA) on the experimental data (Ft. Bragg data). Tables 20 and 21 show the findings for the interactions, the sources of variance most relevant to the question of

Table 20

Ft. Bragg Fakability Study
MANOVA Results
AVOICE Combat-Related Scales

| Type and Name of Scale | Interactions: Fake x Set | Interactions: Fake x Set x Order |
|---|---|---|
| Combat Related Scales: | | |
| Overall | S | NS* |
| Marksman | S | --- |
| Agriculture | S | --- |
| Armor/Cannon | S | --- |
| Vehicle/Equipment Operator | S | --- |
| Outdoors | S | --- |
| Infantry | S | --- |
| Law Enforcement | S | --- |
| Heavy Construction/Combat | S | --- |
| Mechanics | NS | --- |
| Electronics | NS | --- |
| Adventure | S | --- |

Note: S = Significant, $p \leq .01$
NS = Nonsignificant, $p > .01$
* = Marginally significant, $.05 \leq p > .01$
Sample Size is 164

Table 21

Ft. Bragg Fakability Study
MANOVA Results
AVOICE Combat-Support Scales

| Type and Name of Scale | Interactions: Fake x Set | Interactions: Fake x Set x Order |
|---|---|---|
| Combat-Support Scales: | | |
| Overall | S | S |
| Mathematics | S | NS |
| Aesthetics | S | S |
| Leadership | S | S |
| Electronic Communications | S | S |
| Automated Data Processing | S | S |
| Teacher/Counseling | NS | NS |
| Drafting | NS | NS |
| Audiographics | NS | NS |
| Science/Chemical Operations | S | NS |
| Supply Administration | S | NS |
| Office Administration | S | NS |
| Medical Services | NS* | NS |
| Food Services | S | NS* |

Note: S = Significant, $p \leq .01$
NS = Nonsignificant, $p > .01$
* Marginally significant, $.05 \leq p > .01$
Sample Size is 201

whether soldiers can or cannot intentionally distort their responses. As can be seen, nine of the eleven combat-related AVOICE scales are sensitive to intentional distortion, and nine of the twelve combat-support scales are sensitive to intentional distortion. The interaction of fake x set x order is either significant or marginally significant, indicating that order of conditions in which the participant completed the AVOICE also affected the result. Thus, Tables 22 and 23 show mean scores for the various conditions for the particular condition when it was the first administration.

As can be seen in Tables 22 and 23, when told to distort their responses so that they are not likely to be placed in combat-related occupational specialties (MOSs), i.e., instructed to fake non-combat, soldiers tend to decrease their scores on all scales. Scores on 19 of 24 interest scales were lower in fake non-combat as compared to the honest condition. In the fake combat condition, soldiers, in general, increased their combat-related scale scores and decreased their combat-support scale score.

We next examined the extent that the ABLE response validity scales, which had demonstrated they could detect intentional distortion, could be used to adjust AVOICE scale scores for faking combat and faking non-combat. Table 24 shows the mean differences in AVOICE scale scores before and after regressing out ABLE Social Desirability and Poor Impression. Unfortunately, these adjustments have little effect, perhaps because the response validity scales consisted of items from the ABLE and the faking instructions for the ABLE and AVOICE were different. The ABLE faking instructions were fake good and fake bad, whereas, the AVOICE faking instructions were fake combat and fake non-combat.

Another very important finding of the present study is that applicants do not tent to distort their responses to the AVOICE. Tables 22 and 23

Table 22

Fort Bragg AVOICE Combat Scales - Effects of Faking

| AVOICE Combat Scales | Honest* N | Honest* Mean | Honest* S.D. | Fake Combat* N | Fake Combat* Mean | Fake Combat* S.D. | Fake Non-Combat* N | Fake Non-Combat* Mean | Fake Non-Combat* S.D. | Effect Size Honest vs. Combat $X_1 - X_1$ / S.D. Total | Effect Size Honest vs. Non-Combat $X_1 - X_1$ / S.D. Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Marksman | 122 | 18.1 | 4.5 | 58 | 20.2 | 3.9 | 60 | 12.8 | 5.9 | -.49 | 1.06 |
| Agriculture | 124 | 15.0 | 3.8 | 59 | 12.9 | 3.6 | 60 | 15.1 | 4.0 | .56 | -.03 |
| Armor/Cannon | 124 | 14.2 | 5.8 | 59 | 28.9 | 7.6 | 60 | 15.1 | 6.3 | -.73 | 1.53 |
| Vehicle/Equipment | 124 | 28.7 | 6.4 | 59 | 26.6 | 7.9 | 60 | 23.5 | 8.0 | .30 | .75 |
| Outdoors | 123 | 36.0 | 6.1 | 59 | 38.3 | 6.0 | 60 | 25.7 | 10.2 | -.38 | 1.34 |
| Infantry | 123 | 33.5 | 6.8 | 59 | 37.8 | 8.2 | 59 | 20.5 | 8.4 | -.59 | 1.77 |
| Law Enforcement | 124 | 53.3 | 10.8 | 59 | 54.5 | 12.1 | 60 | 42.3 | 12.5 | -.11 | .97 |
| Heavy Construction | 124 | 70.5 | 16.3 | 59 | 68.9 | 15.0 | 59 | 58.7 | 16.4 | .10 | .72 |
| Mechanics | 124 | 50.7 | 12.7 | 59 | 44.6 | 15.2 | 60 | 47.3 | 13.6 | .45 | .26 |
| Electronics | 124 | 58.1 | 18.3 | 59 | 50.3 | 17.3 | 60 | 56.8 | 18.0 | .43 | .07 |
| Adventure | 108 | 37.5 | 4.3 | 56 | 38.1 | 3.7 | 54 | 26.8 | 6.6 | -.15 | 2.06 |

*Values are based on the sample that completed the questionnaire under the condition of interest first.

Table 23

Fort Bragg AVOICE Combat Support Scales - Effects of Faking

| AVOICE Non-Combat Scales | Honest* N | Honest* Mean | Honest* S.D. | Fake Combat* N | Fake Combat* Mean | Fake Combat* S.D. | Fake Non-Combat* N | Fake Non-Combat* Mean | Fake Non-Combat* S.D. | Effect Size Honest vs. Combat $X_1 - X_1$ S.D. Total | Effect Size Honest vs. Non-Combat $X_1 - X_1$ S.D. Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mathematics | 120 | 14.2 | 4.7 | 56 | 11.8 | 4.7 | 59 | 15.6 | 5.0 | .51 | -.29 |
| Aesthetics | 120 | 14.6 | 4.1 | 57 | 12.1 | 4.6 | 59 | 17.1 | 5.5 | .59 | -.54 |
| Leadership | 124 | 22.3 | 4.2 | 59 | 21.5 | 4.1 | 59 | 17.3 | 5.8 | .19 | 1.05 |
| Electronic Communications | 123 | 21.1 | 6.1 | 59 | 21.8 | 7.0 | 60 | 14.2 | 5.6 | -.11 | 1.16 |
| Automated Data Processing | 122 | 20.4 | 6.7 | 58 | 15.5 | 7.2 | 59 | 23.8 | 7.4 | .71 | -.49 |
| Teaching | 124 | 23.8 | 5.7 | 59 | 20.7 | 5.6 | 60 | 21.5 | 5.5 | .64 | .14 |
| Audiographics | 124 | 23.5 | 5.6 | 59 | 18.7 | 6.2 | 60 | 20.7 | 5.6 | .83 | .50 |
| Science | 123 | 28.0 | 8.4 | 59 | 28.0 | 9.2 | 60 | 25.8 | 9.6 | 0 | .25 |
| Supply Administration | 124 | 30.5 | 9.8 | 59 | 26.4 | 9.6 | 60 | 35.3 | 11.9 | .42 | -.46 |
| Office Administration | 123 | 38.5 | 13.3 | 59 | 31.2 | 12.3 | 59 | 49.5 | 17.2 | .56 | -.75 |
| Medical Services | 124 | 67.8 | 18.3 | 59 | 60.4 | 17.5 | 60 | 61.0 | 17.8 | .41 | .37 |
| Food Services | 122 | 38.0 | 10.2 | 59 | 31.0 | 10.6 | 59 | 45.8 | 16.3 | .68 | -.62 |

*Values are based on the sample that completed the questionnaire under the condition of interest first.

Table 24

Effects of Regressing Out Response Validity Scales (Social Desirability and Poor Impression) on Faking Condition AVOICE Scale Scores

| | Fake Good Condition | Fake Bad Condition |
|---|---|---|
| Combat | | |
| Combat-Support | | |

Note: Mean differences are between group comparisons for first testing only. For the Fake Good condition they are [mean (Fake) - mean (Honest)]/SD. For the Fake Bad condition they are [mean (Honest) - mean (Fake)]/SD.

Table 25

Comparison of Ft. Bragg Honest*, Ft. Knox, and MEPS (Applicants)

AVOICE Combat-Related Scales

| AVOICE Combat-Related Scale | Ft. Bragg (Honest)* N | Ft. Bragg (Honest)* Mean | MEPS (Applicants) N | MEPS (Applicants) Mean | Ft. Knox N | Ft. Knox Mean | Total S.D. | Degrees of Freedom | F | P |
|---|---|---|---|---|---|---|---|---|---|---|
| Marksman | 103 | 18.02 | 98 | 17.06 | 187 | 16.00 | 4.79 | 2,385 | 6.26 | .00 |
| Leadership | 103 | 22.54 | 98 | 19.63 | 187 | 20.01 | 4.74 | 2,385 | 12.87 | .00 |
| Electronic Communication | 103 | 20.81 | 98 | 21.69 | 187 | 20.78 | 5.9 | 2,385 | .90 | .41 |
| Armor/Cannon | 103 | 24.14 | 98 | 27.13 | 187 | 22.19 | 6.85 | 2,385 | 18.25 | .00 |
| Vehicle/Equipment Operator | 103 | 28.73 | 98 | 31.22 | 187 | 28.02 | 7.52 | 2,385 | 6.09 | .00 |
| Outdoors | 103 | 36.37 | 98 | 35.57 | 187 | 32.05 | 6.66 | 2,385 | 18.83 | .00 |
| Infantry | 103 | 33.59 | 98 | 33.45 | 187 | 29.20 | 7.51 | 2,385 | 17.45 | .00 |
| Law Enforcement | 103 | 53.83 | 98 | 48.46 | 187 | 48.49 | 11.87 | 2,385 | 7.97 | .00 |
| Heavy Construction/Combat | 103 | 70.75 | 98 | 70.92 | 187 | 65.93 | 17.43 | 2,385 | 3.89 | .02 |
| Realistic | 103 | 3.38 | 98 | 3.10 | 187 | 3.03 | 1.04 | 2,385 | 3.83 | .02 |
| Enterprising | 103 | 3.13 | 98 | 3.01 | 187 | 2.99 | 1.09 | 2,385 | .52 | .60 |
| Adventure | 103 | 37.43 | 98 | 35.51 | 187 | 32.97 | 5.47 | 2,385 | 26.17 | .00 |

*Scores are based on persons who responded to the honest condition first.

Table 26

Comparison of Ft. Bragg Honest*, Ft. Knox, and MEPS (Applicants)

AVOICE Non-Combat-Related Scales

| AVOICE Non-Combat-Related Scale | Ft. Bragg (Honest)* N | Ft. Bragg (Honest)* Mean | MEPS (Applicants) N | MEPS (Applicants) Mean | Ft. Knox N | Ft. Knox Mean | Total S.D. | Degrees of Freedom | F | P |
|---|---|---|---|---|---|---|---|---|---|---|
| Mathematics | 114 | 14.25 | 111 | 13.85 | 231 | 15.02 | 4.49 | 2,453 | 2.90 | .06 |
| Aesthetics | 114 | 14.66 | 111 | 13.03 | 231 | 14.10 | 4.20 | 2,453 | 4.49 | .01 |
| Automated Data Processing | 114 | 20.35 | 111 | 19.01 | 231 | 23.26 | 6.55 | 2,453 | 19.60 | .00 |
| Teacher/Counselor | 114 | 23.88 | 111 | 21.11 | 231 | 22.63 | 5.45 | 2,453 | 7.49 | .00 |
| Drafting | 114 | 22.25 | 111 | 20.93 | 231 | 21.46 | 6.19 | 2,453 | 1.30 | .27 |
| Audiographics | 114 | 23.43 | 111 | 22.22 | 231 | 23.74 | 5.55 | 2,453 | 2.89 | .06 |
| Science/Chemical Operations[†] | | | | | | | | | | |
| Supply Administration[†] | | | | | | | | | | |
| Office Administration[†] | | | | | | | | | | |
| Food Service[†] | | | | | | | | | | |
| Investigative[†] | | | | | | | | | | |
| Conventional[†] | | | | | | | | | | |
| Artistic[†] | | | | | | | | | | |
| Social[†] | | | | | | | | | | |

*Scores are based on persons who responded to the honest condition first.

---

[†]The scoring on these scales is incorrect.

show the mean scores for applicants (MEPS) and the two honest conditions, Ft. Bragg and Ft. Knox. There appears to be no particular pattern to the mean score differences. The applicants score lowest, highest, and in the middle about equal number of times. Though faking or distortion on the interest inventory does not appear to be a problem in an applicant situation, faking or distortion in a draft situation cannot be assessed in the present U.S. situation.

Overall, the AVOICE data from the faking study show that:

1. Soldiers can distort their responses when instructed to do so;
2. The ABLE Social Desirability and Poor Impression scales are not as effective for adjusting AVOICE scale scores in the faking conditions of combat/non-combat as they are for adjusting ABLE content scale scores in the faking good/faking bad conditions; and
3. Faking or distortion does not appear to be a problem in an applicant setting.

### Criterion-Related Validities of the ABLE and AVOICE

We used information from all the above analyses to revise the temperament and interest scales for the trial battery. Those inventories and measures are currently being administered to Army enlisted personnel; the data are not yet gathered that will allow us to evaluate the criterion-related validities of the ABLE and AVOICE.

### Summary

Our overall strategy was to identify temperament, biodata, and interest constructs that were likely to provide incremental validity for predicting on-the-job performance and then to develop reliable measures of those constructs. Our first step was a comprehensive literature review. Those constructs that had demonstrated criterion-related validity in previous research were evaluated by psychologists and the expected true

validities estimated. We then wrote items for scales to measure the most promising constructs. We then administered the scales to soldiers, examined the psychometrics and structural characteristics of the scales, revised the items and scales, and then administered the revised scales to another group of soldiers. We went through this revision/bootstrapping process three times. The reliability of scale scores for the third and last version of the Pilot Trial Battery was thoroughly investigated. We examined internal consistency indices, test-retest reliability, and the reliability of scale scores under different conditions or situations. The results indicate that the internal consistency of the scales is high, the test-retest reliability is high, and though soldiers can distort their responses when instructed to do so, the response validity scales detect such distorting, scores on response validity scales can be used to adjust content scale scores and, perhaps even more important, people in an applicant setting do not tend to distort their responses.

We have again revised the scales that formed the Pilot Trial Battery. They now constitute a part of the Trial Battery which is currently being administered, along with criterion measures, to several thousand soldiers. Soon we will be able to evaluate these predictors according to their criterion-related validity.

## References


Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Guion, R. M., & Gottier. (1965). Validity of personality measures in personnel section. Personnel Psychology, 18, 135-164.

Hough, L. M. (1984). Analyses of 1800 soldiers' scores on Temperament, Biographical, and Interest Measures from the Preliminary Battery. Project A, (Army Research Institute), In-house Progress Review. Minneapolis, MN: Personnel Decisions Research Institute.

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

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. Journal of Applied Psychology, 52, 372-376.

Attachment 1

Problems, Issues, and Results in the

Development of Temperament, Biographical, and

Interest Measures

I. Rationale for including temperament, biographical, and interest predictors

A. Predict on-the-job performance (rather than training)

B. Predict Army-wide criteria (rather than MOS specific criteria)

1. Effort

2. Following Regulations and Orders

3. Leadership

4. Self-Control

II. Strategy and results

A. Identify useful constructs

1. Review literature

2. Develop taxonomies (predictors and criteria)

3. Summarize criterion-related validities reported in literature according to taxonomies (Tables 1, 2, 3)

4. Define tentative list of constructs

5. Obtain expert judgments about expected validities of predictor constructs for Army criteria (Table 4)

6. Select final list of constructs (List 1)

B. Develop Measures

1. Resolve homogeneity/heterogeneity issue

a. Specify components of constructs

b. Write items to tap components (scales)

2. Resolve item and response format issues

a. Temperament Scales (both self-perceptions and biodata)

ABLE (Assessment of Background and Life Experiences)

* Content Scales List 2

* Validity Scales List 2

b. Interest Scales (self-perceptions and biodata)

AVOICE (Army Vocational Interest Career Examination)

* Content Scales List 3

C. Evaluate and revise constructs/scales (issues and results)

1. Evaluate sensitivity of item content

a. psychologist review

b. military review

c. revise items

2. Evaluate psychometric characteristics (Ft. Lewis, Ft. Campbell, Ft. Knox)

a. means

b. standard deviations

c. item response distributions

d. scale score distributions

e. item-total scale score correlations

f. internal consistency of scales (reliability, alpha coefficient)

g. test-retest reliability

h. revise items

3. Evaluate structural characteristics

a. relationships with existing target/defining scales (marker variables)

b. factor analysis

4. Evaluate unique contribution of ABLE and AVOICE to predictor battery (potential for incremental validity)
   a. comparison with ASVAB
   b. comparison with psychomotor measures
   c. comparison with cognitive measures
5. Evaluate response sets
   a. true-false set
   b. inattention to direction, positive versus negative working, of item stem
   c. random responding (response validity scale)
   d. intentional distortion (faking) experiment: 2 x 2 x 2 counter balanced design (Ft. Bragg)

      ABLE: 1. be accepted
            2. be rejected
            3. honest

      AVOICE: 1. like combat-related activities
              2. like non-combat-related activities
              3. honest

      Comparison of mean scores for content and response validity scales:

      1. experimental conditions (Ft. Bragg data)
      2. incumbent data (Ft. Knox data)
      3. applicant data (MEPS data)

6. Evaluate criterion-related validities of ABLE and AVOICE
   a. training criteria
   b. hands-on criteria
   c. Army-wide criteria

# PROBLEMS, ISSUES, AND RESULTS IN THE DEVELOPMENT OF COMPUTERIZED PSYCHOMOTOR MEASURES

Jeffrey J. McHenry and Matthew K. McGue
Personnel Decisions Research Institute

August 1985

Paper presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

## Author Notes

This paper was prepared as part of a symposium on "Expanding the Measurement of Predictor Space for Military Enlisted Jobs," presented at the annual meeting of the American Psychological Association, August, 1985. Each of the papers discusses a different aspect of developing a set of predictor measures for the Army's Project A, an effort designed to improve the selection, classification and utilization of enlisted personnel. 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 is being funded by the U.S. 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 or policies of the Army Research Institute or the Department of the Army.

# Problems, Issues, and Results in the Development of Computerized Psychomotor Measures

Jeffrey J. McHenry Matthew K. McGue

Personnel Decisions Research Institute
Minneapolis, Minnesota

The Armed Services Vocational Aptitude Battery (ASVAB) currently includes no psychomotor tests. Therefore, the psychomotor ability domain was studied carefully in order to determine if psychomotor tests might be useful in improving the predictive utility of the ASVAB. Based on the expert judgment task described in the Peterson (1985) paper, a review of the test validity literature, job observations of first-term soldiers in several MOS, and consideration of testing capabilities, three computerized psychomotor tests were developed and included in a pilot trial test battery.

## Description of Tests

The first test, Target Tracking Test #1, is a pursuit tracking test. For each trial of the test, subjects are shown a path consisting entirely of vertical and horizontal line segments. At the beginning of the path is a target box. Centered in the box is a crosshair. As the trial begins, the target starts to move along the path at a constant, predictable rate of

speed. The subject's task is to keep the crosshair centered within the target at all times. The subject controls crosshair movement via a joy stick mounted on a custom-designed response pedestal (see Figure 1). The maximum crosshair speed, the difference between the speed of the target and the maximum crosshair speed, and the number of turns in the path vary from trial to trial (see Figure 2); however, the total time required for the target to traverse the path is constant. The subject's score on each trial is the mean distance from the center of the crosshair to the center of the target during the course of the trial. The test includes 27 trials. Target Tracking Test #1 was intended to be a measure of precision/steadiness.

The second test, Target Tracking Test #2, is almost identical to Target Tracking Test #1, the only difference being that the subject must use two sliding resistors (instead of a joy stick) to control movement of the crosshair. The sliding resistor mounted on the left half of the response pedestal controls movement of the crosshair in the horizontal plane, while the sliding resistor mounted on the right half of the response pedestal controls movement of the crosshair in the vertical plane. Target Tracking Test #2 was designed to measure multilimb coordination.

The third psychomotor test is the Target Shoot Test. At the beginning of each trial of this test, a target box and crosshair appear on the computer screen. The target might be located anywhere on the screen, but the crosshair always appears in the center of the screen (see Figure 3). The target instantly begins moving about the screen in an unpredictable

manner, frequently changing speed and direction. The subject's task is to use a joystick to move the crosshair onto the center of the target. When this has been accomplished, the subject must press a button on the response pedestal to "fire" at the target. Eventually, if the subject fails to fire at the target, the trial will automatically terminate and a new trial will begin. The parameters varying from trial to trial include: (1) the speed of the target; (2) the maximum speed of the crosshair; and (3) the mean length of each path segment traversed by the target. The subject receives three scores on each trial. The first is whether he/she has hit, missed, or failed to fire at the target. The second is the distance from the center of the crosshair to the center of the target at the time that the subject fires. The third is the time elapsed from the beginning of the trial to the time the subject fires. The second and third scores are obtained only for those trials where the subject fires at the target. The Target Shoot Test was desinged to be a second measure of precision/steadiness, though the test also requires some degree of multilimb coordination.

Major Issues

The analysis of the psychomotor test data focused on the following major issues: (1) how the tests should be scored; (2) how the various test parameters that varied from trial to trial contributed to test difficulty; (3) the effects of machine differences on test scores; and (4) the stability of test scores.

Test scoring. Scoring of Target Tracking Tests #1 and #2 was relatively straightforward. During each trial, the distance from the center of the crosshair to the center of the target was computed approximately 16 times per second, or almost 350 times per trial. These distances were then averaged by the computer, which output only the mean distance for each trial.

The frequency distribution of these mean distance scores proved to be highly positively skewed, the skewness coefficient for some trials being in excess of 5 and 6. Therefore, subjects' mean distance scores for each trial were transformed using the natural logarithm transformation. The overall test score for each subject was then the mean log mean distance across the 27 trials of each test (which shall henceforth be called simply the distance score).

Scoring of the Target Shoot Test was a bit more complicated. Three overall test scores were generated for each subject: (1) the percentage of hits; (2) the mean distance from the center of the crosshair to the center of the target at the time of firing (the distance score); and (3) the mean time elapsed from the start of the trial until firing (the time to fire score). Complications arose because subjects received no distance or time to fire scores on trials where they failed to fire at the target before the time limit for the trial elapsed. This scoring procedure resulted in considerable missing data. Moreover, this missing data occurred primarily on the most difficult items of the test, where only the adept subjects were

able to maneuver the crosshair close enough to the target to fire. Therefore, as a first step in computing overall distance and time to fire scores, the distance and time to fire scores for each trial were standardized. For each subject, the overall distance and time score was then computed by averaging these standardized scores across all trials in which the subject fired at the target.

Effects of Test Parameters

As noted previously, three parameters were used to guide development of Target Tracking Tests #1 and #2: (1) the maximum crosshair speed; (2) the difference between the speed of the target and the maximum crosshair speed; and (3) the number of turns in the path. Table 1 shows that the maximum crosshair speed parameter accounted for over 80% of the within-subject variance in tracking performance in both the Target Tracking Tests. Since the maximum crosshair speed and the target speed were correlated ($r$=.42), this finding means that tracking was much more difficult when the "system" (i.e., target and crosshair) moved faster. In fact, for both tests there was almost a perfect correspondence between the percentage increase in the maximum crosshair speed and tracking accuracy; as the maximum crosshair speed doubled, the tracking error doubled. The speed difference parameter and the three-way interaction between the three parameters were the only other factors with a significant effect on tracking difficulty.

None of the three parameters used to guide development of the Target

Shoot Test accounted for more than 25% of the within-subject variance in any of the test's three dependent measures. Nevertheless, the speed of the target, the maximum speed of the crosshair, and the mean length of each path segment all contributed significantly to the difficulty of this task, as Table 2 shows.

Machine Differences

Past research with psychomotor apparatus tests has shown that subjects' test scores can be significantly affected by differences in mechanical test apparatus. During World War II, for example, psychomotor apparatus tests were an important part of the Army Air Forces' Aircrew Classification Battery. Because the apparatus were extremely sensitive to wear and tear, and because this wear and tear had a major impact on subjects' test scores, test administrators had to invest considerable effort in testing the apparatus, making sure that all parts were operating within tolerance limits, and adjusting or changing worn parts (Melton, 1947). After World War II, the Army, Air Force, and Navy all discontinued apparatus testing, primarily because of the many problems associated with equipment maintenance and the effects of machine differences on test scores (Passey and McClaurin, 1966).

In his paper for this symposium, Rosse (1985) described the hardware configuration assembled for computer test data collection. He also noted how final calibration of the response pedestal was accomplished via a

special software calibration routine. It was hoped that this calibration program would control for differences between joy sticks, dials, sliding resistors, buttons, etc., and eliminate any machine effects on overall test scores.

To determine the effectiveness of the calibration program, a multivariate analysis of variance (MANOVA) was executed. The independent variable was the computer testing station. Six testing stations were used in total, and approximately 35 subjects were tested on each station (each subject was tested at one station). The dependent variables were 19 selected test scores from the 10 computerized tests. These included four scores from the psychomotor tests--the distance scores from Target Tracking Tests #1 and #2 and the distance and time to fire scores from the Target Shoot Test. Results of the MANOVA are shown in Table 3. Machine differences had no effect on test scores (MANOVA likelihood ratio=.99, $p$=.50). Even for individual test scores, none of the 19 $p$ values is less than .05. Looking specifically at the psychomotor tests, only the time to fire score from the Target Shoot Test approaches statistical significance ($p$=.09); $p$ values for the remaining three scores all exceed .40. This is most encouraging, since one would suspect that the variability among joy sticks and sliding resistors used in the psychomotor tests would be the most likely cause of any machine effects on test scores.

Stability of Test Scores

One of the major concerns in psychomotor testing is the stability of

test scores. Research on motor skills learning shows that subjects' performance on new motor tasks tends to improve very quickly at first, then more slowly over time (Singer, 1980). Task proficiency may not reach an asymptote, however, until the task has been performed thousands of times (Crossman, 1959; Seibel, 1964). In studying this learning phenomenon, many researchers have found that the abilities contributing to task proficiency change as the task is performed and practiced (Adams, 1953, 1957; Fleishman, 1957; Fleishman and Rich, 1963; Hinrichs, 1970). For example, spatial ability and verbal ability tend to be good predictors of proficiency during the initial stages of motor task learning, but they are not particularly useful predictors of task proficiency after subjects have received extensive practice. If the ability requirements of a motor task change with practice, it would not be surprising to find that correlations between scores on the first few trials of the task are not highly correlated with scores on trials occurring after extensive practice.

To study the effects of practice on subjects' test scores, an experiment was conducted (see Figure 4). Subjects who were tested on the computer battery at Time 1 returned two weeks later for testing at Time 2. When they returned, they were assigned to one of two experimental groups. The retest group completed the entire computer battery a second time. The practice group was retested on just five tests: Target Tracking Tests #1 and #2, the Target Shoot Test, the Choice Reaction Time Test, and the Cannon Shoot Test. Prior to retesting, the practice group received two rounds of practice on each of the five tests. Each practice round consisted of 15

trials of each of the Target Tracking Tests, 20 trials of the Target Shoot Test, 15 trials of the Choice Reaction Time Test, and 24 trials of the Cannon Shoot Test. If they wished, subjects were allowed to take a short break between the two practice rounds or between the second practice round and the retest, but these breaks never exceeded five minutes.

Data from the practice experiment were used to answer two important questions: (1) Would the performance of the practice group improve more than the performance of the retest group on these five tasks? and (2) Would the test-retest stability of the tests be greater for the retest group than for the practice group? Table 4 provides the answers to these questions.

First, Table 4 shows that, at best, the test performance of the practice group improved only slightly more than the retest group. The difference between the gain scores for the two groups reached statistical significance only for the distance score for Target Tracking Test #2 and for the time to fire score for the Target Shoot Test. These data suggest that the practice intervention was not a particularly strong one. It should be noted, though, that on many tests subjects' performance actually deteriorated from Time 1 to Time 2. The average gain score for the two groups across the six dependent measures was only .23 standard deviations. (Ignoring the time to fire score from the Target Shoot Test reduces the average gain score to .09 standard deviations.) This suggests either that the tasks used in these tests are resistant to practice effects, or that

performance on these tasks reaches a maximum level of proficiency after only a few trials.

Next, Table 4 shows that the test-retest stability for all six dependent measures was greater for the retest group than the practice group. (While the difference between the stability coefficients for the two groups was statistically significant for only one of the six dependent measures, the test was not very powerful; statistical significance required a difference of approximately .40 between the two stabilities.) Closer inspection of the data shows that the stability coefficients for the two groups were very nearly equal for the three "distance" dependent measures. Thus, it appears that the rank-ordering of subjects' performance on psychomotor tests is not greatly affected by practice.

Of the six dependent measures, the only one which appears to be greatly affected by practice is the time to fire score on the Target Shoot Test. Even though their distance scores improved only marginally relative to those of the retest group, during Time 2 testing the practice group fired at the target much earlier in the trial than the retest group. It is impossible to say whether this behavior was a result of an improvement in tracking skills or a result of a change in firing strategy (e.g., firing as soon as the crosshair was inside of or near the target instead of waiting until the crosshair was perfectly centered inside the target). The data also show that the time to fire score was highly unstable for the practice group ($r_{xx}$=.13). The firing strategy of these subjects during the retest session

was totally unrelated to their firing strategy during the initial test session.

Another method for studying practice effects is to examine the correlations between items or parts within a test. This was done for Target Tracking Tests #1 and #2. Each test was divided into three parts corresponding to test items 1-9, 10-18, and 19-27. A distance score was then computed for each of the three parts. Table 5 shows the intercorrelations among the three part scores for both tests for both Time 1 and Time 2. (Time 2 data were taken from the retest group only; the practice group's data were not included.) If the ability requirements of the tracking task were changing due to practice during the course of the test, one would expect to find that the correlation between items 1-9 and items 19-27 would be lower than either of the two correlations involving items 10-18. This did not occur. While there is a slight tendency for the correlation between items 10-18 and items 19-27 to be the highest of the three intercorrelations, the difference between the highest and lowest correlation within each test averages only .05. Other data show that the Spearman-Brown corrected split-half reliability of both tests is .97, suggesting that all of the items within each test are measuring the same underlying ability.

In summary, scores from computerized psychomotor tests appear to be quite stable over a two-week period. While practice does effect some test scores, practice apparently has little impact on tracking accuracy.

Implications for Future Test Development

The data from the three psychomotor tests indicate that all three tests meet acceptable psychometric standards of reliability. They are relatively free of error variance due to machine or practice effects. The data also shed some light on the parameters affecting the difficulty of each of the three tests.

The data suggest that certain refinements are in order before proceeding with additional testing. The positively skewed frequency distributions of the distance scores from Target Tracking Tests #1 and #2 indicate that both tests are too easy, and that they are not differentiating well among high ability individuals. To make the tests more difficult, the target and crosshair should be made to move faster in future versions. Given their high internal consistency reliabilities and the lack of practice effects, it would also appear that the tests can be shortened considerably without jeopardizing reliability. If the tests were cut by a third to 18 items each, for example, the Spearman-Brown corrected split-half reliability should still be approximately .95.

More dramatic changes are required for the Target Shoot Test. The chief problem is to reduce the amount of missing data. The only way to accomplish this is to reduce the difficulty of the items, so that subjects have an opportunity to fire at all (or almost all) targets. Unfortunately,

this may drastically alter some of the characteristics of the test. If the items are too easy, all subjects will likely be able to achieve low distance scores, making the time to fire score perhaps the only meaningful dependent measure. Currently, though, the time to fire score is very susceptible to practice effects. Both the distance score and the time to fire score will therefore have to be monitored carefully in subsequent versions of this test to determine if the characteristics of these scores (e.g., test-retest stability, practice effects, etc.) change in an undesirable manner.

References

Adams, J. A. (1953). *The prediction of performance at advanced stages of training on a complex psychomotor task* (Research Bulletin 53-49). Lackland Air Force Base, TX: U. S. Air Force Human Resources Research Center.

Adams, J. A. (1957). The relationship between certain measures of ability and the acquisition of a psychomotor criterion response. *Journal of General Psychology*, *56*, 121-134.

Crossman, E. R. F. W. (1959). A theory of the acquisition of speed-skill. *Ergonomics*, *2*, 153-166.

Fleishman, E. A. (1957). A comparative study of aptitude patterns in unskilled and skilled psychomotor performances. *Journal of Applied Psychology*, *41*, 263-272.

Fleishman, E. A., and Rich, S. (1963). Role of kinesthetic and spatial-visual abilities in perceptual-motor learning. *Journal of Experimental Psychology*, *66*, 6-11.

Hinrichs, J. R. (1970). Ability correlates in learning a psychomotor task. *Journal of Applied Psychology*, *54*, 56-64.

Melton, A. W. (Ed.) (1947). *Apparatus tests* (Army Air Forces Aviation Psychology Program Research Report No. 4). Washington, DC: U. S. Government Printing Office.

Passey, G. E., and McLaurin, W. A. (1966). *Perceptual-psychomotor tests in aircrew selection: Historical review and advanced concepts* (PRL-TR-66-4). Lackland Air Force Base, TX: U. S. Air Force Personnel Research Laboratory.

Peterson, N. G. (1985). *Overall strategy and methods for expanding the measured predictor space.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Rosse, R. L. (1985). *Advantages and problems with using portable computers for personnel management.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Seibel, R. (1964). Levels of practice, learning curves, and system values for human performance on complex tasks. *Human Factors, 6,* 293-298.

Singer, R. N. (1980). *Motor learning and human performance: An application to motor skills and movement behaviors* (3rd ed.). New York: Macmillan.

TABLE 1

Proportion of Within-Subject Variance Accounted for by the Three Target Tracking Test Parameters

| Parameter | Proportion of Variance Accounted for in Target Tracking Test #1 | Proportion of Variance Accounted for in Target Tracking Test #2 |
|---|---|---|
| Number of Turns (A) | 0.6% | 1.0% |
| Maximum Crosshair Speed (B) | 80.8% | 82.9% |
| Difference between Maximum Crosshair Speed and Target Speed (C) | 6.7% | 6.8% |
| A x B | 0.2% | 0.7% |
| A x C | 1.0% | 2.6% |
| B x C | 2.6% | 2.6% |
| A x B x C | 8.0% | 6.1% |

TABLE 2

Proportion of Within-Subject Variance Accounted for by the Three Target Shoot Test Parameters

| Parameter | Proportion of Variance Accounted for in the Time to Fire Score | Proportion of Variance Accounted for in the Distance Score |
|---|---|---|
| Maximum Crosshair Speed | 8.5% | 17.9% |
| Target Speed | 8.4% | 22.5% |
| Mean Segment Length | 3.3% | 6.3% |

TABLE 3

The Effects of Machine Differences on Computer Test Scores[a]

| Computer Test Score | F | p[b] |
|---|---|---|
| Simple Reaction Time - Mean Reaction Time | 1.59 | .16 |
| Choice Reaction Time - Mean Reaction Time | .52 | .76 |
| Perceptual Speed & Accuracy - Percent Correct | 1.18 | .32 |
| Perceptual Speed & Accuracy - Mean Reaction Time | .56 | .73 |
| Perceptual Speed & Accuracy - Slope | .84 | .53 |
| Perceptual Speed & Accuracy - Intercept | .85 | .52 |
| Target Identification - Percent Correct | 1.67 | .14 |
| Target Identification - Mean Reaction Time | .93 | .46 |
| Short Term Memory - Percent Correct | .11 | .99 |
| Short Term Memory - Mean Reaction Time | .95 | .45 |
| Short Term Memory - Slope | 1.13 | .34 |
| Short Term Memory - Intercept | .64 | .67 |
| Cannon Shoot - Time Score | 2.14 | .06 |
| Number Memory - Percent Correct | .56 | .73 |
| Number Memory - Mean Response Time | 1.55 | .17 |
| Target Tracking Test #1 - Mean Log Distance | .62 | .69 |
| Target Tracking Test #2 - Mean Log Distance | .86 | .51 |
| Target Shoot - Time to Fire | 1.91 | .09 |
| Target Shoot - Mean Distance | 1.01 | .41 |

a. MANOVA likelihood ratio=.99, p=.50 for these 19 test scores.
b. df = 5,200 for all 19 test scores.

TABLE 4

Gain Scores and Reliabilities
for Retest and Practice Groups[a]

| Test | Test Score | Group | Gain Score[b] | F for Gain Scores | Stability | Z for Stabilities[c] |
|---|---|---|---|---|---|---|
| Target Tracking #1 | Distance | Retest | .07 | 4.11 | .68 | .46 |
| | | Practice | .33 | | .64 | |
| Target Tracking #2 | Distance | Retest | -.09 | 7.79* | .77 | .16 |
| | | Practice | .21 | | .76 | |
| Target Shoot Test | Distance | Retest | .21 | .08 | .58 | .88 |
| | | Practice | .26 | | .48 | |
| Target Shoot Test | Time to Fire | Retest | -.18 | 176.19* | .47 | 2.56* |
| | | Practice | 2.47 | | .13 | |
| Choice Reaction Time | Mean RT | Retest | -.36 | .73 | .56 | 1.64 |
| | | Practice | -.43 | | .36 | |
| Cannon Shoot Test | Time Score | Retest | .34 | 5.72 | .66 | 1.50 |
| | | Practice | -.11 | | .51 | |

a. Inferential statistics significant at $p<.01$ are denoted with an asterisk.
b. Gain scores are effect size estimates. Signs were reflected so that a positive gain score denotes "improvement" from Time 1 to Time 2.
c. Given the sizes of the retest and practice samples, statistical significance will not be attained until the difference between the two stabilities reaches approximately .40.

TABLE 5

Intercorrelations among Items 1-9, Items 10-18, and Items 19-27 of Target Tracking Tests #1 and #2

Target Tracking Test #1

| | Time 1 | | | | Time 2 | | |
|---|---|---|---|---|---|---|---|
| | Items 1-9 | Items 10-18 | Items 19-27 | | Items 1-9 | Items 10-18 | Items 19-27 |
| Items 1-9 | | | | Items 1-9 | | | |
| Items 10-18 | .87 | | | Items 10-18 | .91 | | |
| Items 19-27 | .80 | .87 | | Items 19-27 | .92 | .92 | |

Target Tracking Test #2

| | Time 1 | | | | Time 2 | | |
|---|---|---|---|---|---|---|---|
| | Items 1-9 | Items 10-18 | Items 19-27 | | Items 1-9 | Items 10-18 | Items 19-27 |
| Items 1-9 | | | | Items 1-9 | | | |
| Items 10-18 | .83 | | | Items 10-18 | .86 | | |
| Items 19-27 | .85 | .89 | | Items 19-27 | .85 | .91 | |

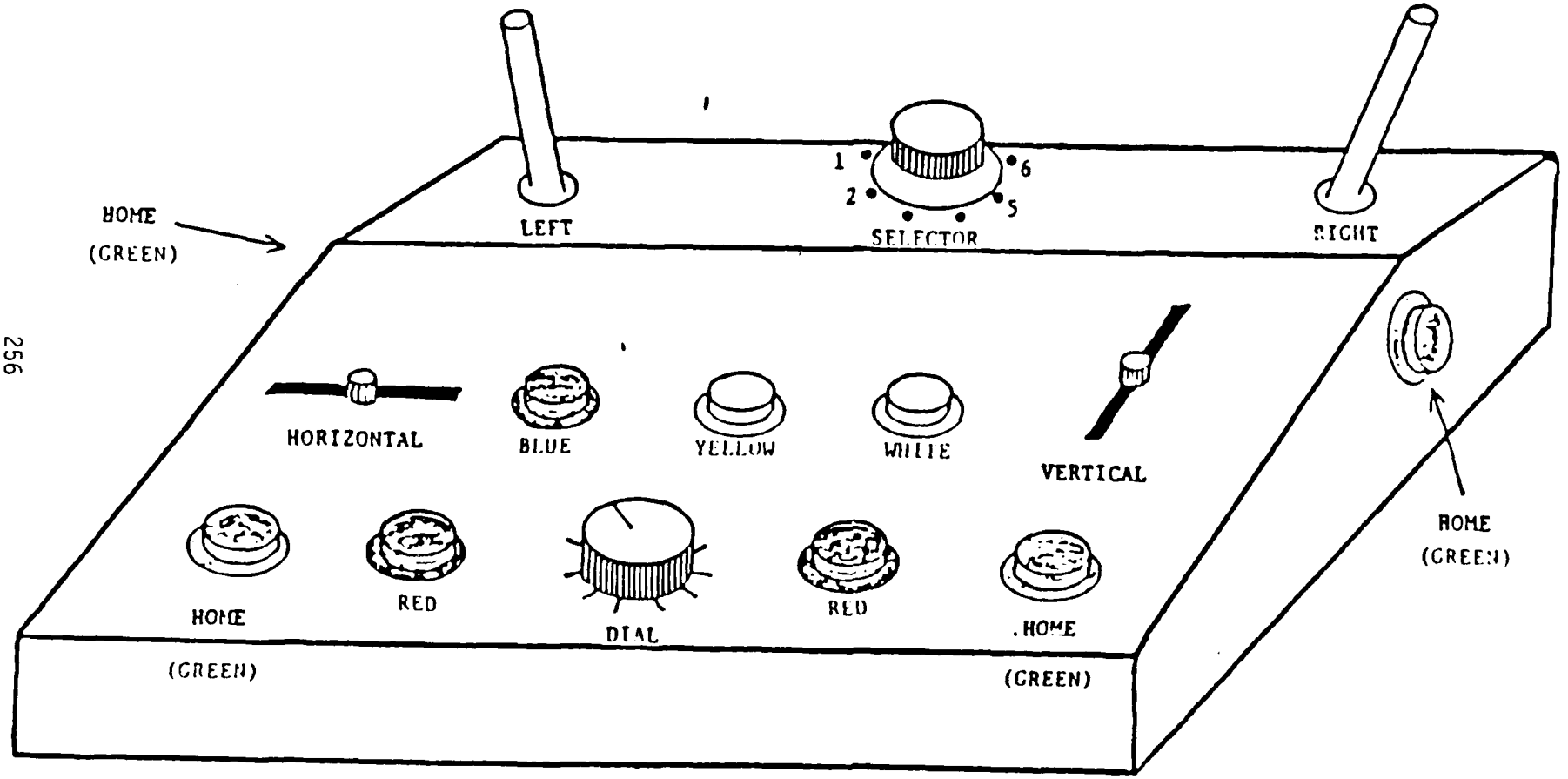



FIGURE 1. Custom-designed response pedestal

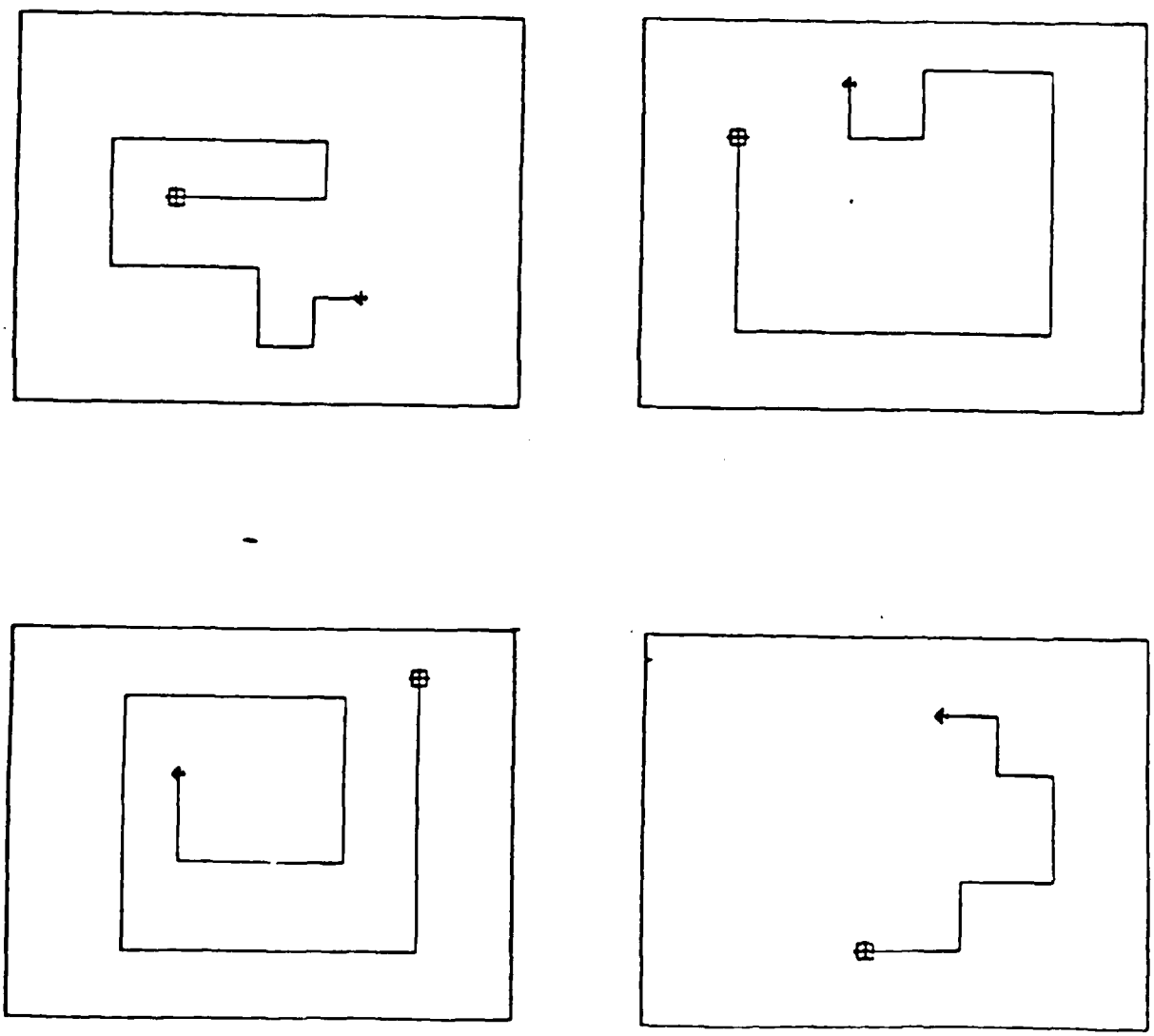
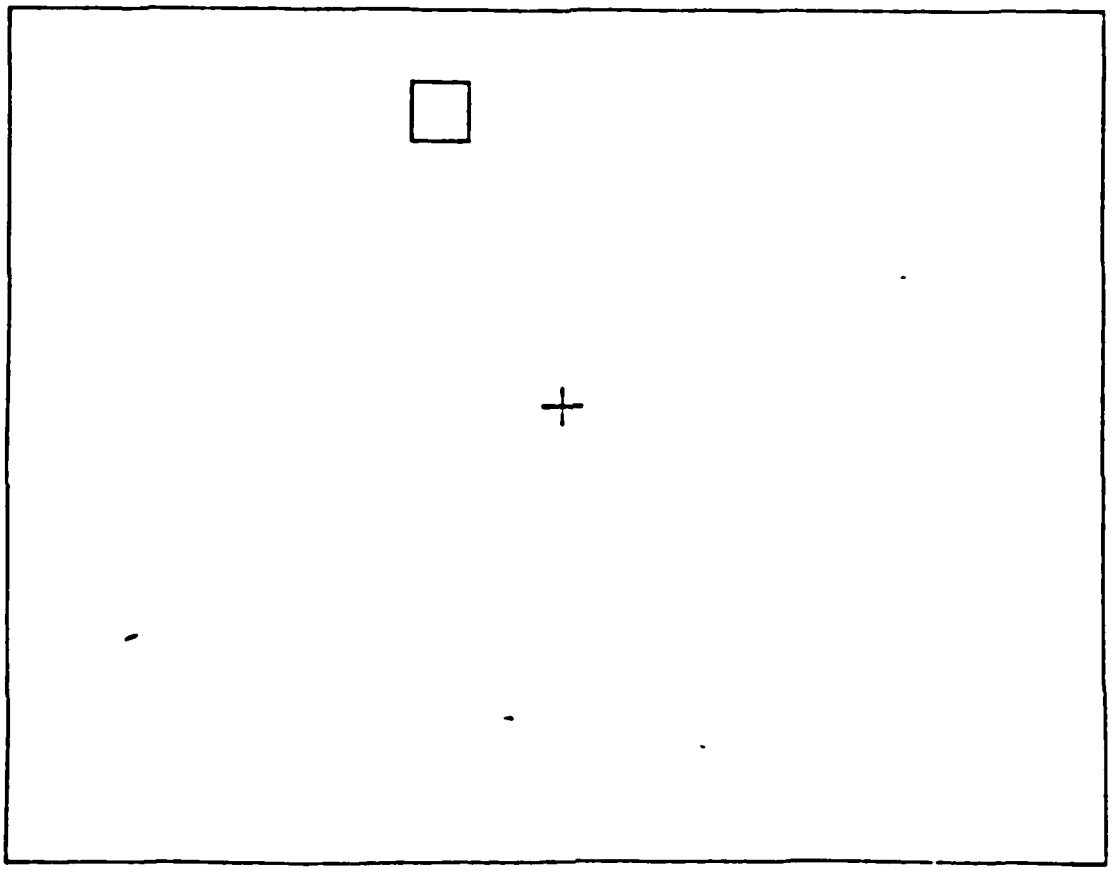
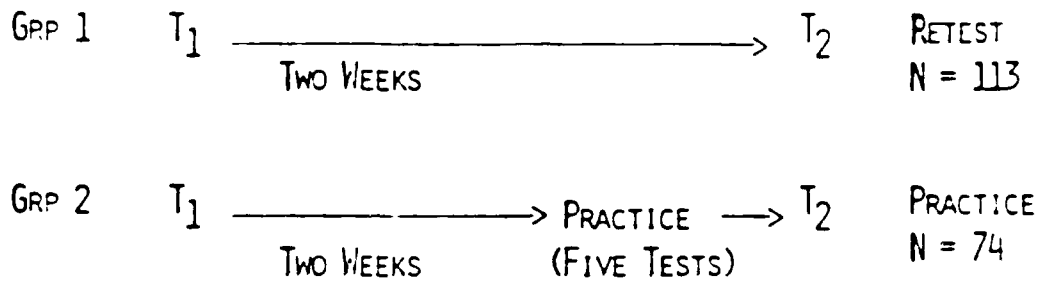


FIGURE 2. Sample Target Shoot Test Items



FIGUPE 3. Sample Target Shoot Test item




ANOVA

| Source | DF |
|---|---|
| Group | A-1 |
| Subjects (Grp) | (B-1)A |
| Time | C-1 |
| Time x Grp | (C-1)(A-1) |
| Time x Subj (Grp) | (C-1)(B-1)A |

Practice Effect = Significant Time x Grp Interaction

Stability = $r_{T_1T_2}$ for Group 1

FIGURE 4. Experimental design of the practice effects study

# MEASUREMENT OF TEST BATTERY VALUE FOR SELECTION AND CLASSIFICATION

Donald H. McLaughlin
American Institutes for Research

August 1985

Presented at the Annual Meeting of the American Psychological Association, Los Angeles, California

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

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

PROBLEM

Last year I reported to APA on our project's efforts to identify optimal composites for ASVAB forms 8, 9, and 10, in terms of differential validity as well as overall absolute validity (McLaughlin, 1984). This work is continuing, and in the course of the next year, we must again select optimal composites, and we must also select the most valuable battery of new measures from a large set being developed for enlisted personnel selection and classification. For this purpose, an overall framework for evaluation of the value of a battery is needed, to replace ad hoc, intuitive weighting of competing criteria. In this paper, I describe progress in the development of such a framework.

The value of a prediction instrument for personnel selection and classification is the average increment in expected utility of performance that can be gained by assigning individuals to jobs in accordance with predictions based on the instrument, as compared to random assignment. Lord (1952) described a framework for measuring this value, although, based on an assumption of normally distributed payoff functions, he noted (and I concur) that the "necessary expressions at present available for the multiple integrals are too cumbersome to be of practical use."

In the course of exploring these multiple integrals I have obtained a few results, however. These pertain to the approximate estimation of the <u>value</u> of a prediction instrument, defined as:

(1) $$V = (1/N)\,(\mathrm{Sum}_{ij}(d_{ij}\hat{u}_{ij}) - \mathrm{Sum}_{j}(N_j\,\mathrm{ave}_{i}\hat{u}_{ij}))$$

where $\hat{u}_{ij}$ is the estimated utility of performance for individual i when assigned to job j,

$d_{ij}$ is 1 or 0, taking on the value 1 if the optimal assignment for individual i based on the instrument is to job j; and

N and $N_j$ are the total number of individuals to be assigned and the number to be assigned to job j, respectively.

This definition encompasses both absolute validity, the gain from selecting the best of a set of applicants, and differential validity, the gain from assigning the selected applicants to jobs that match their qualifications. The subtracted term in the definition can be dropped if we make the non-restricting assumption that the utility predictions are scaled to have means of zero. With this assumption, the expected utility for random assignment is zero, providing the baseline for V. Furthermore, we assume that utility of performance is scaled to be in equal utility units across different jobs. The scaling of the $\hat{u}_{ij}$, the expected utility estimates, is further governed by the validity of the measures available for predicting $\hat{u}_{ij}$. To simplify the presentation, I will use a single notation, $\hat{u}_{ij}$, to indicate expected utility, noting, like Cronbach & Gleser (1965), that it is a product of various factors; in particular, of the relative value of performance increments in different jobs and the predictability of those performance increments.

To combine both selection and classification into one framework, we merely add one more classification category to the M job classifications, j=M+1, and define $\hat{u}_{i,M+1}$ to be identically zero. The choice of zero for the value of this constant is immaterial, because adding a constant to the expected utility of all applicants for a particular category does not change the expected gain relative to random assignment.

If we consider that applicants' expected utilities are distributed multivariately, as $f(\hat{U})$, then we can further specify:

$$(2) \qquad V = \sum_{j=1}^{M} \int_{R_j} \hat{u}_{ij} \, f(\hat{U}) \, d\hat{U},$$

where $R_j$ is the region in which j is the best job to assign to individuals, and there are M different jobs.

As Lord (1952) pointed out, the regions can be defined by the half-hyperplanes that separate them, and in fact, the problem of finding the optimum assignment can be translated into finding the single point at which these half-hyperplanes intersect.

Although the general problem is somewhat difficult, I have obtained a few results from approximations to this integral for the multivariate normal distribution. These are shown in Table 1. These values of V assume that $v_j$ is the standard deviation of $u_{.j}$. The first four rows of Table 1 are for the two-category case. First, the general two-category value of V is the product of the standard deviation of the difference, $u_{.1} - u_{.2}$, times the value of the normal density function at the point at which the split between the two categories is in the specified ratio, p:1-p. The latter factor achieves its maximum, 1/sqrt(2pi), when p=.5. This measure can be written in terms of the Classification Efficiency Index which Paul Horst proposed a while back (Horst, 1954) for application to the unconstrained classification problem. It is the square root of CE, times a multiplier which, for the two category case, only varies with the level of imbalance in the requirements for the two categories. The major substantive issue I will address here is the extent to which that fact generalizes beyond the two-category problem. If there is sufficient generality, then we can use an adaptation of CE, referred to as $H^*_2$ in Table 1, to compare alternative prediction batteries and alternative sets of composites based on a prediction battery for use in constrained selection and classification situations.

The zero expected utility for the rejected catagory is frequently irrelevant to the evaluation of a test battery, so we have divided the entries in rows 2 and 3 by $p_s$, taking the average gain over selected applicants.

The remaining entry in Table 1 refers to special cases for more than two categories. They refer to cases in which the covariance matrix of estimates is a scalar multiple of the identity matrix and in which equal numbers of applicants are to be assigned to each category. The 3-category solution is exactly 3/2 times the two-category solution, but the relationship to number of categories is not that simple. Based on an approximation to the inverse of the normal distribution function, it is possible to approximate $H^*_M$ for balanced assignment with uniform validities and uncorrelated predictors by a sum of terms, as shown in Table 1. This expression is valid for any number of categories, and the graph of V as a function of number of categories is shown in Figure 1.

Table 1. Explicit Expressions for the Value Measure for Special Cases for Multivariate Normally Expected Utilities

---

Two Categories (M=2)

General Expression: $H_2^* = \text{SQRT}( D_{12}^2 ) f(p_1)$

$= \text{SQRT}( v_1^2 - 2 r_{12} v_1 v_2 + v_2^2 ) f(p_1)$

Even Split, Equal Validities: $H_2^* = v \text{ SQRT}((1-r_{12})/\text{pi})$

Selection Only ($v_2=0$) $H_2^* = v f(p_s)/ p_s$

Even Split, Selection Only: $H_2^* = v / (p_s \text{ SQRT}( 2 \text{ pi }))$

Equal Validities and Requirements, and Zero Intercorrelations

General Approximation:

$$H_M^* = 4.7983 \text{ M } ( 1/(\text{M}+.1443424) - \sum_{k=0}^{M-1} ( (-1)^k \binom{M-1}{k} / (k+1.1443424) ) ) v$$

Values for Small M:

| | |
|---|---|
| 2 | .5645 v |
| 3 | .8467 v |
| 4 | 1.0298 v |
| 5 | 1.1633 v |
| 6 | 1.2676 v |

---

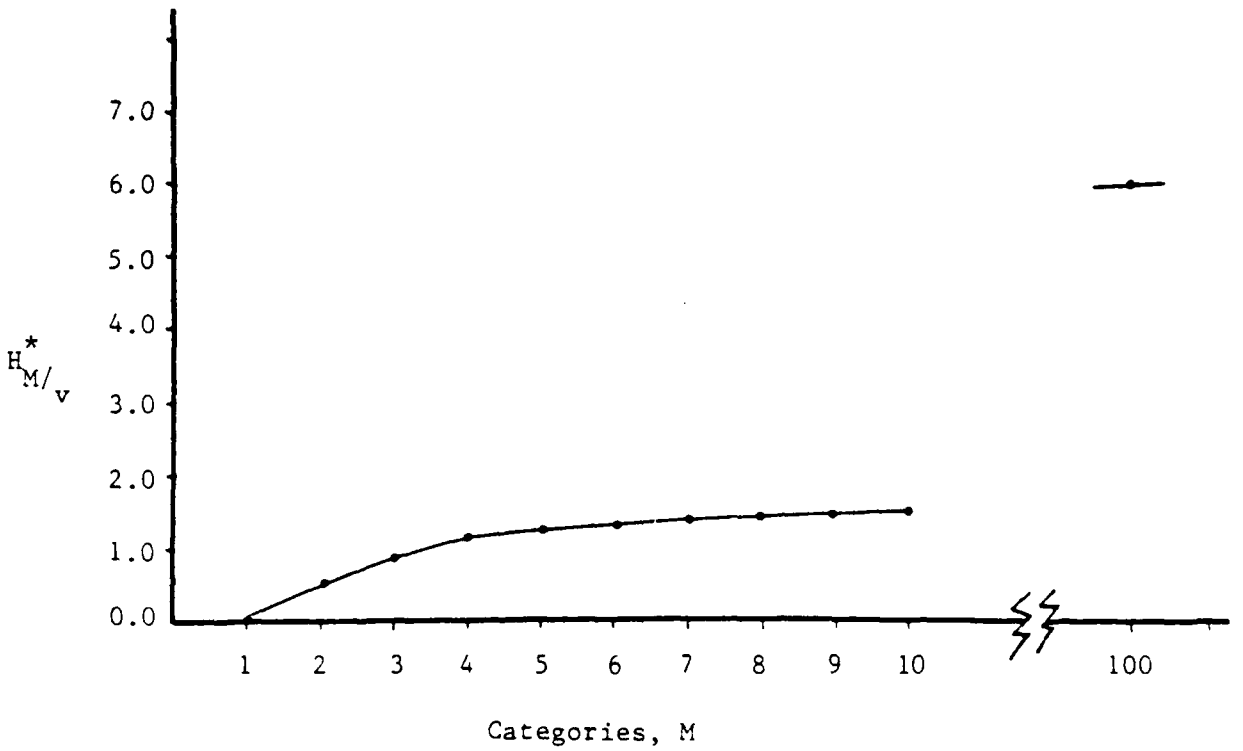



Figure 1. Relationship of H* to number of categories.

The ordinate in Figure 1 is the expected gain from using a battery with validities (i.e., expected utility standard deviations) all equal to 1, uncorrelated estimates, and balanced assignments. The gain would be multiplied by the validity level, for validities other than 1.

The meager results indicated in Table 1 provide a starting point for development of a measure of battery-validity for constrained selection and classification. We need to identify a general expression for $H^*_M$, allowing for unequal validities, nonzero intercorrelations, and unbalanced assignment requirements. The expression should assign the same value to functionally identical batteries, and it should fit the results presented in Table 1. One choice is the following:

$$(3) \qquad H^*_M = c_M \, \mathrm{SQRT}\left( \mathrm{Sum}_{jk} \left( D^2_{jk} f^2_{jk} (p_j + p_k) \right) \right),$$

where j and k range over pairs of categories;
$p_j$ and $p_k$ are the proportions of individuals required for categories j and k;
$D^2$ is the variance of the difference in estimators for categories j and k;
f is the normal density function at the point where the split matches the ratio of $p_j$ to $p_k$; and
$c_M$ is a constant related to the overall number of categories.

Note that this index will be zero unless there is some pair of categories for which $D^2_{jk}$ is greater than zero; in the latter case, $H^*_M$ will be greater than zero.

The value of $D^2_{jk}$ is simple to compute for the two-category case, as can be seen in Table 1. For more categories, however, the value of any $D^2_{jk}$ is increased due to the special restriction of range that occurs due to assignment of many individuals to other "third" categories, as shown in Figure 2.

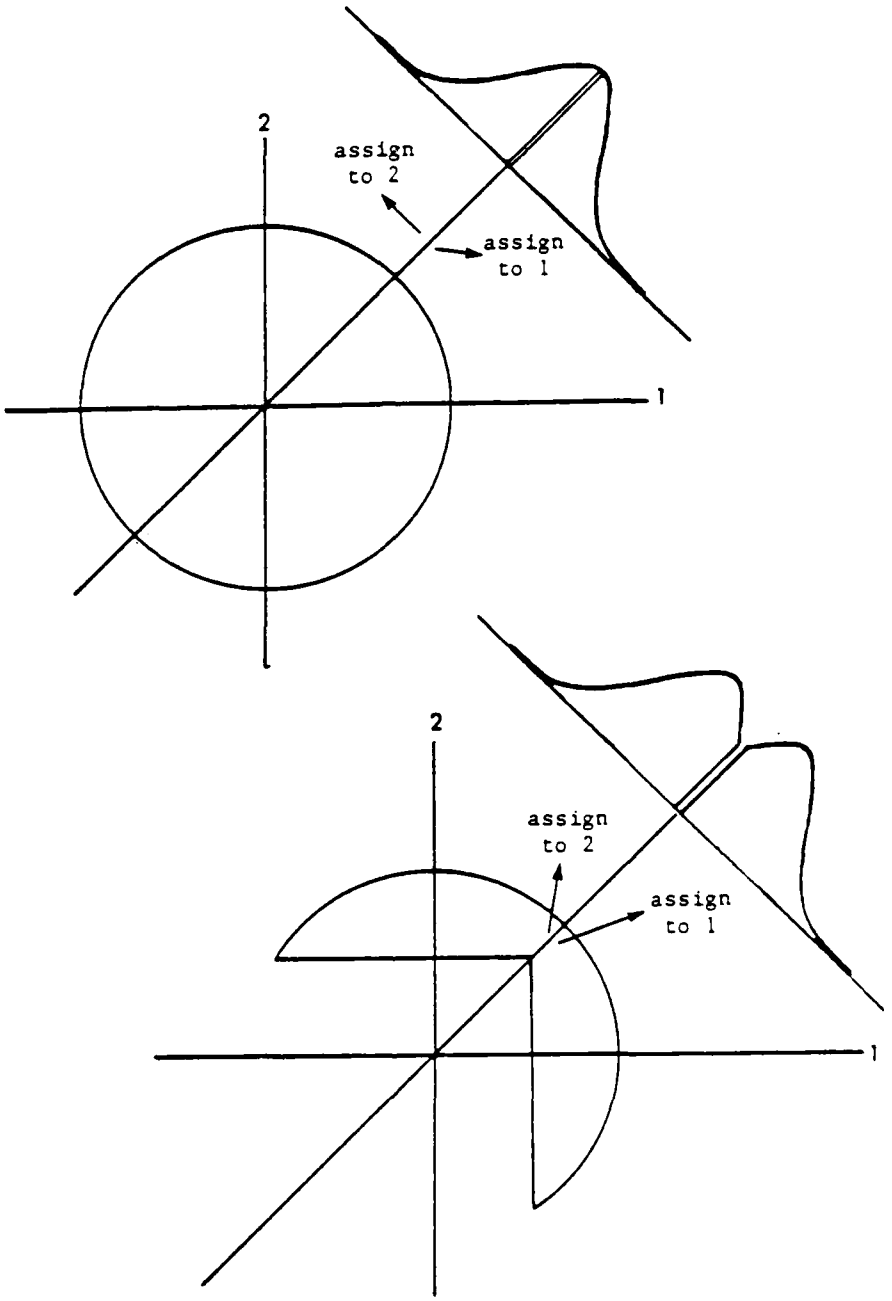



Figure 2. Effective "separation" of two categories due to assignment of many individuals to a third category.

METHOD

The general method was to compute sample estimates of V for selection and classification of individuals distributed according to known multivariate normal distributions, and then to compare these estimates to $H^*_M$, in order to identify:

(1) how $H^*_M$ performs for balanced constrained assignments;

(2) whether $H^*_M$ for unbalanced assignments can be approximately factored into its value for balanced assignments times a function purely of the level of imbalance; and if so,

(3) an approximate form of the function of imbalance in requirements for different categories.

Eight sets of 5,000 cases were generated, as described in Table 2. Deviations from the nominal values of the parameters were small. Over all data sets, the average observation was .008 greater than the nominal value, the standard deviations were .004 greater than the nominal values, and the correlations were .005 greater than the nominal values. Each case was a point in a five-dimensional normal variate space. Each set of 5,000 was divided into 10 files of 500 cases, and the optimal assignment was determined for each file, for a variety of requirement alternatives. Because the cost of each optimization was roughly $2.00 to $6.00, we elected to use only 10 file replicates of each condition. The stability of the results supports this choice.

Optimizations were performed for the sets of requirements shown in Table 3, not in a factorial design crossed with data sets, but to address particular specific questions.

Table 2. Simulated Data Sets

| | Validities | Intercorrelations |
|---|---|---|
| Data Set 1 | .5,.5,.5,.5, 0 | .4,.4,.4,.4,.4,.4 |
| Data Set 2 | .5,.5,.5,.5, 0 | .6,.6,.6,.6,.6,.6 |
| Data Set 3 | .5,.5,.5,.5, 0 | .8,.8,.8,.8,.8,.8 |
| Data Set 4 | .5,.5,.5,.5, 0 | .8,.8,.8, 0, 0, 0 |
| Data Set 5 | .8,.6,.4,.2, 0 | .4,.4,.4,.4,.4,.4 |
| Data Set 6 | .8,.6,.4,.2, 0 | .8,.8,.8,.8,.8,.8 |
| Data Set 7 | .7,.7,.7,.7, 0 | .4,.4,.4,.4,.4,.4 |
| Data Set 8 | .7,.7,.7,.7, 0 | .8,.8,.8,.8,.8,.8 |

$$\begin{pmatrix} v_1 & r_1 & r_2 & r_4 & 0 \\ & v_2 & r_3 & r_5 & 0 \\ & & v_3 & r_6 & 0 \\ & & & v_4 & 0 \\ & & & & 0 \end{pmatrix}$$

Selected Categories

Rejected Category

Table 3. Alternative Category Size Requirements Simulated

| Category 0 (Rejected) | Category 1 | Category 2 | Category 3 | Category 4 |
|---|---|---|---|---|
| 0 | 25% | 25% | 25% | 25% |
| 0 | 12% | 21% | 29% | 38% |
| 0 | 5% | 5% | 45% | 45% |
| 52% | 12% | 12% | 12% | 12% |
| 52% | 6% | 10% | 14% | 18% |
| 52% | 2.4% | 2.4% | 21.6% | 21.6% |
| 52% | 0 | 16% | 16% | 16% |
| 52% | 0 | 0 | 24% | 24% |
| 52% | 0 | 0 | 0 | 48% |

The method of optimization was based on Lord's framework: because the shape of the optimal regions is known, finding the optimum "merely" involves finding the point of intersection of all the boundary half-hyperplanes. A simple iterative procedure was used to move from a starting estimate of this point toward the optimum, where each step consisted of making assignments according to the trial boundary intersection point and recording the resulting excesses and deficits in each category. A new trial boundary intersection point was selected to move in the direction of maximally reducing the deviations from the requirements. With intelligently selected starting points, the solution was found in roughly a dozen iterations in nearly every case.

## RESULTS

The first step is to verify the applicability of $H_M^*$ for balanced assignments. The results in Table 4, based on 4-category classifications (100 % selection), indicate that $H_M^*$ is nearly as accurate for balanced constrained assignments as for the unconstrained assignments that were the basis for Horst's derivation. The unconstrained values were obtained from the first iteration of the optimization, using a starting value of zero on all differences (i.e., setting all regions to the same size). Sample errors of estimate were computed across the ten replications of 500 case assignments in each condition. Because the same seed was used for the random number generator in each set of 5,000 cases, the differences were, in fact, much stabler than indicated by the sample errors of estimate.

To summarize the results in Table 4, the expected utility gain from use of the battery is directly proportional to the square root of the difference between the average of the squared validities of the estimates and the average of the covariances of the estimates, in balanced constrained classification as well as in unconstrained classification. In fact, the loss due to the constraint that the numbers in each category be exactly equal was less than 1%. Overall, $H_M^*$ closely matched the obtained values of V when validities and intercorrelations were homogeneous but slightly overestimated V when they varied.

Table 4

Relationship Between $H_4^*$ and Computed Gains Due to Validity, for Balanced Assignment Into Four Categories

| Validities for Separate Categories | Intercorrelations of Estimators | $H^*$ | Mean Computed Gain (SD): Unconstrained Assignment | Mean Computed Gain (SD): Constrained Assignment, V | $H^* - V$ |
|---|---|---|---|---|---|
| .5 | all .4 | .403 | .407 | .406 | -.003 |
| .5 | all .6 | .334 | .334 | .334 | .000 |
| .5 | all .8 | .243 | .241 | .240 | +.003 |
| .7 | all .4 | .569 | .569 | .569 | .000 |
| .7 | all .8 | .336 | .336 | .336 | .000 |
| .8,.6,.4,.2 | all .4 | .475 | .470 | .457 | +.018 |
| .8,.6,.4,.2 | all .8 | .354 | .348 | .337 | +.017 |
| .5 | .8,.8,.8,0,0,0 | .411 | .393 | .374 | +.037 |

The effect on V of variation in the validity level of a uniform set of validities is clear. As can be seen in Figure 3, the ratio of V for pairs of conditions that differ only in that the separate validities were .5 in one case and .7 in the other case was 1.4, with little if any systematic error. This result is not sensitive to imbalances in the requirements for different jobs. The formula, $H_M^*$, captures this regularity.

The interactions between (1) imbalances in requirements and (2) mean predictor intercorrelations affect the value of a battery, as shown in Table 5. Five different levels of balance and imbalance are shown, for both cases including selection and cases consisting purely of classification of selectees. The values for the cases involving both selection and classification are adjusted (by dividing by .48) to ignore the utility component for the rejectees, which would be zero.

There is an apparent regularity in these value estimates. The values for intermediate values of imbalance can be well approximated from the endpoints of complete balance and complete imbalance, but only using empirically determined multipliers. As noted in the last row of Table 5, these multipliers are roughly approximated by the measure of imbalance provided by information theory, but the theoretical basis for such an approximation, if it exists, is not clear.

Three other results are apparent in Table 5. First, the fact that the same multipliers work for both the classification only and the classification plus selection situations indicates the possibility of separating these two phases, computing the added "differential" validity added by a battery beyond its "absolute" validity. Unfortunately, this fact does not appear to generalize to cases in which the validities are not uniform. Second, the value of .753 for the combined classification and selection is fairly well fit by $H_5^*$, which for this case was .761. The other point to be observed in Table 5 is that the ratio of the values between r=.4 and r=.8, which according to the formula should be sqrt(3), is roughly approximated by sqrt(3), but differently for the cases involving and not involving selection.

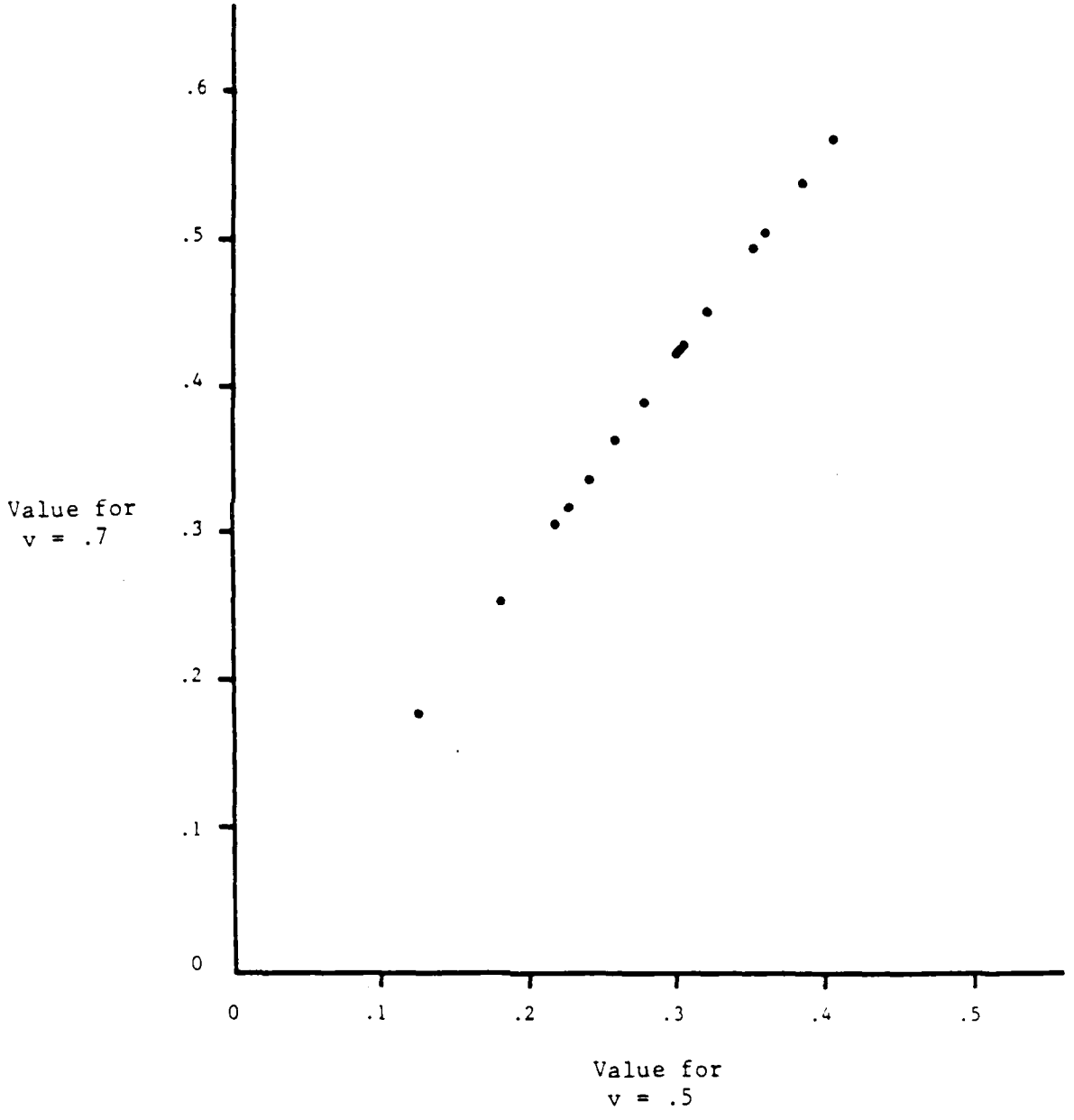



Figure 3. Effect of subtest validity level (v) on test battery value.

Table 5. Relationship of Imbalances in Requirements to Gains in Expected Utility, for Uniform Validities and Intercorrelations. (Formula Estimates are Given in Parentheses)

| | Percentage Breakdown for Four Categories | | | | |
|---|---|---|---|---|---|
| | 25,25,25,25 | 38,29,21,12 | 45,45, 5, 5 | 50,50,0,0 | 100,0,0,0 |
| Classification only | | | | | |
| r=.4 | .406 | .384 (.385) | .302 (.301) | .218 (.217) | .000 |
| r=.8 | .240 | .228 (.228) | .180 (.178) | .126 (.128) | .000 |
| Classification after Selecting 48% of Applicants | | | | | |
| r=.4 | .753 | .736 (.736) | .666 (.669) | .600 (.602) | .428 |
| r=.8 | .635 | .625 (.624) | .581 (.582) | .540 (.539) | .428 |
| Multiplier: | 1.000 | .948 | .742 | .534 | .000 |
| Relative Uncertainty: | 1.000 | .948 | .734 | .500 | .000 |

Note: Computations were for data sets with estimator validities (i.e., expected utility standard deviations) equa to .5. Formula estimates in parentheses are linear interpo'ations between the endpoints in each row of the table, using the multipliers shown.

Next, we examine the effect of variance among intercorrelations on the value of a battery. Table 6 shows the incremental utility gains using batteries with validities of .5 but with four different configurations of intercorrelations. The important comparison is between the first two rows, both of which have mean intercorrelations of .4. The value is larger for the uniform intercorrelations than for the nonuniform intercorrelations. Rough interpolation would indicate that the pattern of intercorrelations of .8,.8,.8,.0,.0,.0 is about equivalent to uniform intercorrelations of .5, not of .4. The implication of this is that if test development effort is to be expended to decrease composite intercorrelations, that effort should be spread across subtest areas, other things equal.

Finally, we examine cases in which the validities vary between the categories. In Table 7, values are shown for various combinations of imbalance and validity variation. The ratios of these values to corresponding values for uniform validities are shown in parentheses. Generally, variance in validities increases the value of the tests, especially when the intercorrelations are high. The fact that a few of the comparisons do favor the uniform validities in this Table is due to the setting of greatest requirements in categories with the lowest validities. For example, in the next to last row, only 2.4% of the applicants are assigned to the job with validity of .8, while 21.6% of the applicants are assigned to the job with validity of .2.

The overall pattern in Table 7 is complex, different for the cases involving selection and not involving selection. The values appear to be related to (1) the average level of intercorrelations, (2) the information measure of imbalance, and (3) the covariance (across categories) between validities and size requirements.

Table 6. Effects of Variation of Intercorrelations of Estimators on Computed Gains

| | Percentage Breakdown into Categories | |
|---|---|---|
| | (25,25,25,25) | (12,12,12,12, 52% rejected) |
| All r=.4 | .406 | .753 |
| r=.8,.8,.8,0,0,0 | .374 | .720 |
| All r=.6 | .334 | .706 |
| All r=.8 | .240 | .635 |

Note: Computations were performed on data sets for which the validities of the estimators were .5. The zero intercorrelations in the second row are all with the same estimator.

Table 7. Effects of Variation in Validities of Separate Estimators on Computed Gains

| Percentages in Categories with Different Validities | | | | | Intercorrelations | | |
|---|---|---|---|---|---|---|---|
| | | | | | All = .4 | All = .8 | All = 1.0 |
| .8 | .6 | .4 | .2 | .0 | | | |
| 25 | 25 | 25 | 25 | 0 | .457 (+13%) | .337 (+40%) | .207 (+inf) |
| 12 | 21 | 29 | 38 | 0 | .395 ( +3%) | .296 (+30%) | .190 (+inf) |
| 5 | 5 | 45 | 45 | 0 | .268 (-11%) | .201 (+12%) | .135 (+inf) |
| 45 | 45 | 5 | 5 | 0 | .400 ( -2%) | .275 (+14%) | .135 (+inf) |
| 12 | 12 | 12 | 12 | 52 | .766 ( +2%) | .673 ( +6%) | .535 (+25%) |
| 6 | 10 | 14 | 18 | 52 | .653 (-11%) | .573 ( -8%) | .464 ( +9%) |
| 2.4 | 2.4 | 21.6 | 21.6 | 52 | .492 (-26%) | .440 (-24%) | .370 (-14%) |
| 21.6 | 21.6 | 2.4 | 2.4 | 52 | .855 (+28%) | .765 (+32%) | .617 (+44%) |

Notes: Comparisons to gains computed for the vector (.5,.5,.5,.5,0) of validities are shown in parentheses. "inf" refers to the fact that the expected gain for uniform validities in these cases is zero.

## SUMMARY

These relations represent aspects of an overall framework for evaluating the combination of differential and absolute validity of a test battery in terms of the added value of making personnel assignments in line with the battery. Although the derivation of an exact expression of the required multiple integrals may be impossible, these results indicate that an approximation formula may be sufficient to make test development decisions. If the approximation is close, then even if it leads to selection of a less than optimal battery, the difference between the selected and optimal batteries will be extremely small.

The major problems that remain to resolve in refinement of the formula are (1) the size of the "separating effect" between two categories caused by the assignment of large numbers of individuals to other categories and (2) the specific form of the effects of imbalances in requirements for different occupational categories on the measure of test battery value.

## REFERENCES


Cronbach, L. J. & Gleser, G. C. (1965) Psychological tests and personnel decisions. Univ. of Illinois Press, Urbana.

Horst, P. (1954) A technique for the development of a differential prediction battery. Psychol. Monog, 68 (9), 1-30.

Lord, F. M. (1952) Notes on a problem of multiple classification. Psychometrika, 17, 297-304.

McLaughlin, D. H. (1984). Differential validity of the ASVAB for job classification. Paper presented at APA Convention, Toronto.

# OVERALL STRATEGY AND METHODS FOR EXPANDING THE MEASURED PREDICTOR SPACE

Norman G. Peterson
Personnel Decisions Research Institute

August 1985

Paper presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

Author Notes

This paper was prepared as part of a symposium on "Expanding the Measurement of Predictor Space for Military Enlisted Jobs," presented at the annual meeting of the American Psychological Association, August, 1985. Each of the papers discusses a different aspect of developing a set of predictor measures for the Army's Project A, an effort designed to improve the selection, classification and utilization of enlisted personnel. 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 is being funded by the U.S. 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 or policies of the Army Research Institute or the Department of the Army.

## Introduction

Our applied problem is to expand the presently measured predictor space for the ultimate purpose of accurately selecting persons for the U.S. Army and appropriately classifying those persons into jobs or Military Occupational Specialties (MOS). In this paper, I describe the strategy we have adopted, the thinking behind the strategy, and the progress that has been made following our strategy. First, a bit more discussion of the research problem is in order.

Presently, the U.S. Army has a lot of jobs and hires, almost exclusively, inexperienced and untrained persons to fill those jobs. As obvious as those facts are, they are still the overriding facts to be addressed. One implication of these facts is that a highly varied set of individual differences' variables must be put into use to stand 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 obvious 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 that do not have the training and experience to immediately begin performing their assigned jobs. This is so

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).

## Theoretical Approach

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 (a) identifying the major, relatively independent domains or types of individual differences' constructs that existed; (b) selecting measures of constructs within each domain that met a number of psychometric and pragmatic criteria, and (c) 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 ASVAB. If the new measures were not relatively independent of 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 much more straightforwardly implemented.

Figure 1 presents an illustrative, construct-oriented model and is presented in order to represent the model in abstract. Note that both the criterion and predictor space are depicted. (A great deal of the work of Project A is devoted to the criterion side, and we, on the predictor side, have taken advantage of the information coming from those efforts as they have become available.)

If this illustrative model were to be developed and tested with data, then the network of relationships on the predictor side, the criterion side, and between the two could be confirmed, disconfirmed, and/or modified. It goes without saying, but I will say it anyway, that the development of such models must be

| 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 | | | | | | | | |
| | Sociability | . | . | . | M | M | M | L | L |

FIGURE 1. Illustrative Construct-Oriented Model

*Denotes expected strength of relationship, High, Medium, Low.

done very carefully and conservatively, and subjected frequently to reality testing. We have kept this firmly in mind. Note, however, that 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 or not 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 columns. We have found it very useful to employ this hierarchical notion, since 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. (See Peterson and Bownas, 1982, for further discussion of this type of a *model*.)

## Research Objectives

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 entry into the Army, successful soldier performance in general and in classifying persons into MOS where they will be most successful, with special emphasis on attributes not tapped by current pre-induction 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 pilot tests and to the concurrent and pre ictive validation samples.
4. Estimate and evaluate the reliability of the new pre-induction 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 pre-induction measures and current pre-induction measures.
6. Determine the degree to which the validity of new pre-induction measures generalizes across MOS, i.e., proves useful for predicting measures of successful soldier performance across quite different MOS and, conversely, the degree to which the measures are useful for classification or the differential prediction of success across MOS.
7. Determine the extent to which new pre-induction measures increase the accuracy of prediction of success and the accuracy of classification into MOS over and above the levels of accuracy reached by current pre-induction measures.

## Research Design and Organization

Design. To achieve these objectives, we have followed the design depicted in Figure 2. There are fifteen sub-tasks in our actual research plan, each tied to one or more of the activities

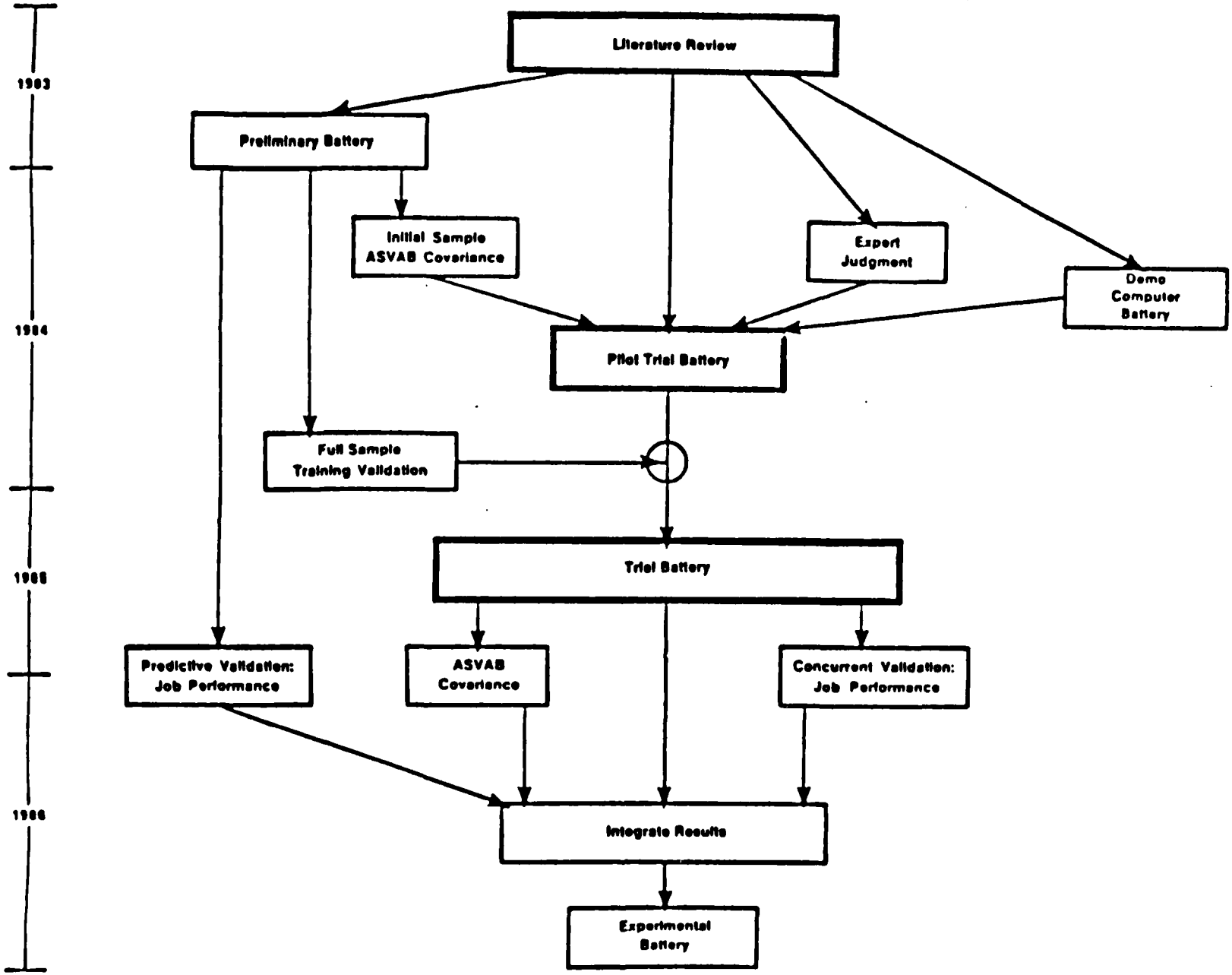



FIGURE 2. Flow Chart of Predictor Measure Development Activities of Project A

or products depicted in Figure 2. At this point, we have completed the activities leading up to the development of the Trial Battery and are now collecting data in the field with that set of measures.

The next year will see the completion of the "Predictive Validation: Job Performance" (on the Preliminary Battery), investigation of the covariance of the Trial Battery with ASVAB, and the "Concurrent Validation: Job Performance" of the Trial Battery using the data presently being collected.

There are several things that we feel are noteworthy about the design. First, there are five test batteries 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 utilized early in the project in order 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. Wing, Peterson, and Hoffman (1984) reported on the expert judgment findings last year.

Thirdly, 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 2. Thus, we are able to benefit from the advantage of both types of designs, i.e., 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 ourselves 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 for short. The final team concerned itself with psychomotor variables and was labeled the "psychomotor" team or sometimes the "computerized" team, since all the measures developed by that team were computer-administered. The activities and fruits of the labor of those teams are reported on in some detail by the various other papers in this symposium: Toquam et al. (1985),

McHenry and McGue (1985), Hough et al. (1985), and, Rosse and Peterson (1985).

## Progress

One gauge of our progress is the degree to which we have accomplished the seven research objectives earlier presented. A second way of looking at progress is to describe the evolution of our predictor model.

### Achievement of Research Objectives

1. Identify "best bet" measures--this objective has been met, i.e., we were able to sift through a mountain of literature, translating the information onto a common form that enabled later evaluation of constructs and measures in terms of several psychomotor 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 which 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 sample and

not-so-small sample tryouts of these measures as they were developed, as is documented by the other papers. The Trial Battery presently being administered in the field is the tangible product of meeting this objective. We have at least one more iteration to go through yet--the modification of the Trial Battery based on analysis of this summer's data. (We are targeted to collect data on about 11,000 soldiers in nineteen MOS.)

3. Developing 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 squandered. We think we have developed and implemented effective methods for getting maximum quality and quantity of data out of our data collection efforts, and the favorable results we have so far achieved in terms of completeness and usefulness of data are due in large part to the attention paid to this objective.
4. Estimate reliability and vulnerability of measures--this objective has also been largely accomplished, and we can happily report that analyses to data indicate that the new measures are psychometrically sound and acceptably invulnerable to the various sources of measurement problems--or we have come up with some ways to adjust for such effects. However, it is safe to say that further, more specifically targeted, research would be very useful in this

area.

5. Determine the interrelationships between the new measures and current pre-induction 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. and 7. Determine the level of prediction of soldier performance, classification efficiency, and incremental validity of the new measures--the jury is still out on these questions since we are just now collecting the data that will enable us to address these objectives.

## Evolution of the Predictor Model

We began our research with a general kind of model, very much like the one presented in Peterson and Bownas (1982). That is, we conceived of the predictor space as divided into several domains with major, relatively independent constructs falling into each domain. At this early point in the research, we were most concerned with thinking about the predictor space in a way guided by past research that would also provide "handles," if you will, for us to approach our particular applied problem. Thus, we formed the domain teams earlier mentioned to be responsible for broad pieces of this predictor space model.

The domain teams reviewed the literature and identified the most promising constructs and wrote definitions and other descriptive materials for each of 53 identified constructs. These constructs were input to the expert judgment process (Wing, et al, 1984), and analysis of those 35 judges' data gave rise to the hierarchical model shown in Figure 3, reproduced from the Wing paper. Remember, this model represents the covariances of the predictors to one another based on their judged relationships to U.S. Army enlisted criterion dimensions.

We then focused our effort on developing measures that would comprehensively cover this predictor space, when those new measures would be combined with the ASVAB measures.

Figure 4 shows the names of the measures included in the Pilot Trial Battery that were developed to measure the model of the predictor space. (Note--ABLE is the Assessment of Background and Life Experiences, and AVOICE is the Army Vocational Interest Career Examination, each containing a number of distinct scales. Both of these instruments are covered in the Hough et al. paper.) These measures were carefully scrutinized, based on a fairly large scale pilot test, and several of the papers in this symposium report on those analyses.

Because of testing time constraints, not all of the measures in the Pilot Trial Battery could be carried forward. We, therefore, reduced the number of measures about 33%. This reduction problem was approached via the model, and one goal was

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

FIGURE 3. Hierarchical Map of Predictor Space

| PILOT TRIAL BATTERY | CLUSTERS | FACTORS |
|---|---|---|
| ASVAB | A. Verbal Ability/ General Intelligence | |
| Reasoning 1 and 2 | B. Reasoning | |
| Number Memory (c) | C. Number Ability | COGNITIVE ABILITIES |
| Perceptual Speed and Accuracy (c)<br>Target Identification (c) | H. Perceptual Speed and Accuracy | |
| AVOICE | U. Investigative Interests | |
| Short Term Memory (c) | J. Memory | |
| Reasoning 1 and 2 | F. Closure | |
| Assembling Objects<br>Object Rotation<br>Shapes<br>Mazes<br>Path<br>Orientation 1, 2, and 3 | E. Visualization/Spatial | VISUALIZATION/ SPATIAL |
| Simple Reaction Time (c)<br>Choice Reaction Time (c) | G. Mental Information Processing | INFORMATION PROCESSING |
| ASVAB | L. Mechanical Comprehension | MECHANICAL |
| AVOICE | M. Realistic vs. Artistic Interests | |
| Target Tracking 1 (c)<br>Target Shoot (c) | I. Steadiness/Precision | |
| Target Tracking 2 (c)<br>Target Shoot (c) | D. Coordination | PSYCHOMOTOR |
| -- | K. Dexterity | |
| ABLE/AVOICE | Q. Sociability | |
| AVOICE | R. Enterprising Interest | SOCIAL SKILLS |
| ABLE | T. Athletic Abilities/Energy | VIGOR |
| ABLE | S. Dominance/Self-esteem | |
| ABLE | N. Traditional Values/Conventionality/Non-delinquency | MOTIVATION/ STABILITY |
| ABLE | O. Work Orientation/Locus of Control | |
| ABLE | P. Cooperation/Emotional Stability | |
| Cannon Shoot (c) | Movement Judgment | |

(c) = Computerized Measures

FIGURE 4. Pilot Trial Battery Measures of the Modeled Predictor Space

to retain as comprehensive coverage as possible of the predictor space. I will not describe that decision process in detail here, but I will say that the model shown in Figure 4, but, in more detail, was in the forefront of that entire process.

The resulting "Trial Battery" takes a bit less than four hours to administer and is generally described in Figures 5 and 6. As mentioned earlier, this battery is now being administered to about 11,000 soldiers in nineteen MOS, and job performance criterion data are also being collected. With these data in hand, we should be able to make some fairly definitive tests, and, no doubt, modifications of our model.

| | |
|---|---|
| PERCEPTUAL/PSYCHOMOTOR (COMPUTER) | 34% |
| COGNITIVE PAPER AND PENCIL | 50% |
| NON-COGNITIVE PAPER AND PENCIL | 16% |

FIGURE 5. Percent of Trial Battery Testing Time Devoted to Each Area

| Motor/Perceptual | Aptitude | Bio-Data/Temperament |
|---|---|---|
| Perceptual Speed and Accuracy | Assembling Objects | Assessment of Background and Life |
| 1 Hand Tracking | Object Rotation | Experiences (ABLE) |
| Target Identification | Maze | |
| 2 Hand Tracking | Reasoning | Interest |
| Cannon Shoot | Orientation/Rotation | Army Vocational Interest Career |
| Target Shoot | Orientation/Map | Examination (AVOICE) |
| Reaction Time | | |
| Short Term Memory | | |
| Number Memory | | |



FIGURE 6. COMPOSITION OF THE PREDICTOR BATTERY

## References


Hough, L.M., Barge, B.N., Houston, J.S., McGue, M.K., & Kamp, J.D. (1985). Problems, issues and results in the development of temperament, biographical, and interest measures. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

McHenry, J.J., & McGue, M.K. (1985). Problems, issues, and results in the development of computerized psychomotor measures. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Peterson, N.G., & Bownas, D.A. Skill, task structure, and performance acquisition. In Marvin D. Dunnette and Edwin A. Fleishman (Eds), Human performance and productivity (Vol. I). Hillsdale, N.J.: Lawrence Erlbaum Associates, 1982.

Rosse, R.L., & Peterson, N.G. (1985). Advantages and problems with using portable computers for personnel measurement. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Toquam, J.L., Dunnette, M.D., Corpe, V., McHenry, J.J., Keyes, M.A., McGue, M.K., Houston, J.S., Russell, T.L., & Hanson, M.A. (1985). Development of cognitive/perceptual measures: Supplementing the ASVAB. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

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

# ADVANTAGES AND PROBLEMS WITH USING PORTABLE COMPUTERS FOR PERSONNEL MEASUREMENT

Rodney L. Rosse and Norman Peterson
Personnel Decisions Research Institute

August 1985

Paper presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

## Author Notes

This paper was prepared as part of a symposium on "Expanding the Measurement of Predictor Space for Military Enlisted Jobs," presented at the annual meeting of the American Psychological Association, August, 1985. Each of the papers discusses a different aspect of developing a set of predictor measures for the Army's Project A, an effort designed to improve the selection, classification and utilization of enlisted personnel. 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 is being funded by the U.S. 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 or policies of the Army Research Institute or the Department of the Army.

## Introduction

"History repeats itself" is an adage that probably does not apply to the advances of microprocessor developments. Given the frantic rate of development, it is difficult to imagine that circumstances could ever again occur in just the way that they did at the outset of this effort in the Fall of 1983. It would seem, however, that any 1986 project might be enhanced by consideration of both the occasional wisdom and sometime folly of our beginning efforts.

To say that the first step should be to decide what is to be done may seem trite. In fact, at least for us, the goals to be accomplished were far from obvious and may have remained beyond our vision except for the valuable help obtained through visits to several research centers doing advanced work in computerized testing: (1) Air Force Human Resources Laboratory at Brooks Air Force Base, Texas, (2) Army Research Institute Field Unit at Fort Rucker, Alabama, (3) Naval Aerospace Medical Research Laboratory, Pensacola, Florida, and (4) Army Research Institute Field Unit at Fort Knox, Kentucky. Experimental testing projects using computers at these sites had already produced impressive developments which stimulated the ideas of the project at hand and have continued to influence our work.

In this paper, we focus primarily on the process we followed and some problems we encountered in hardware and software acquisition and development for the purpose of developing

experimental tests of abilities that could best be administered via microprocessors. As mentioned in other papers in this symposium, the overall goal of the larger research project was to develop and evaluate new predictors in terms of additional prediction/classification accuracy for Army enlisted jobs.

## Hardware Acquisition and Development

Much of the detail of the planned products was yet to evolve at the point of acquisition of the first six machines so that we had to focus upon more general objectives. It was clear that we wished to accomplish several things which were either difficult or impossible to accomplish with paper-and-pencil testing. Specifically, we required the ability to have a very high degree of precision in stimulus presentation and a high degree of control of respondent behavior. Variables were specifically expected to include precision in timing of stimulus presentation and response speed.

In addition, we know that pursuing applied research efforts, even those on the relatively large scale of this project, automatically invokes limitations upon what research questions may be considered. It was intended from the outset that a product for a particular kind of usage be produced in a particular manner, within a particular budget, and on a limited schedule that permitted cooperation of varied resources and groups.

Microprocessor

The choice of which microprocessor to use for the preliminary development was not obvious or straightforward. The arrays of available microcomputer devices were, at the time, in transition from earlier machines which used the first popular microprocessor chips (ie., 8080 or Z-80) into a newer variety of options created by the influence of IBM's entry into the market with their "PC" employing the newer 8088, 8086-7 chips. With the newer machines came more flexible operating systems (e.g., DOS 1 or DOS 2).

A computer designed for portable use was deemed to be a highly desirable characteristic because the machines were to be frequently disassembled, carried or shipped to new locations, and reassembled by personnel with minimal experience in computer hardware. Such portable machines had been on the market only briefly at the time so that little reported experience with them was available.

We acquired six machines made by Compaq (TM) which appeared to suit the need. They were among the "newer" types of machines which used a variation of the MS-DOS operating system. They were equipped with standard game adapters which permitted the analog inputs from "off-the-shelf" joysticks and boolean input from game button switches.

The choice was specifically made to avoid using color in the visual displays for at least two reasons: (1) the certainty of

individual differences in color vision among military recruits, and (2) dread of the prospects of attempting to calibrate video colors for standardization of presentation. Accordingly, we precluded the possibility of directly investigating the value of stimulus effects in color presentation. The machines that were chosen had green on black screens which, it was reasoned, would produce results that could be generalizable to any other monochrome display with relatively little risk.

The graphics capability of the Compaq microcomputer proved to be minimally acceptable for the applications which were to come. In graphics mode, the pixels (or dots) on the screen are organized into 200 rows and 640 columns. To form a figure, one sets the pixels by coordinates to form the desired pattern. More recently, several computers of the "personal" computer type are offering 400 rows with 640 columns which should provide improved resolution.

Very accurate timing of events occurring in the testing process was essential. Initially, timing was accomplished by two means: (1) accessing the calendar clock that is available in any machine which uses MS-DOS (or the variations of MS-DOS that are sold under computer tradenames), and (2) use of calibrated software loops. Without delving too far into technical details, those two options eventually presented some difficulties because of time consumption in the process of obtaining the time. For

instance, the computer CPU often had to be tied up with timing events when other work required being done in the timed interval.

A wonderful solution to the timing problem eventually presented itself in what the computer people call a "real-time-clock" which can be added to the "IBM-type" microcomputers for as little as $50. It operates with a small battery and it is ordinarily sold for personal computers so that the correct date and time will be maintained without resetting when the computer is turned off.

With appropriate software, the "real-time-clock" device allows the timing of events accurately to the nearest 1/1000-th of a second with neglible loss of computer time in the reading. (The sub-program used in our projects will read the time in approximately 1/3000-th of a second.)

Peripheral Devices for Response Acquisition: Response Pedestal

The initial choices in the hardware configuration for a "testing station" proved satisfactory for the "stimulus side", i.e., the controlled presentation to the subject. The standard keyboard and the "off-the-shelf" joysticks were hopelessly inadequate for the "response side." Computer keyboards leave much to be desired as response acquisition devices--particularly when response latency is a variable of interest. Preliminary trials using, say, the "D" and "L" keys of the keyboard for "true" and "false" responses to items was troublesome with naive subjects. Intricate training was required to avoid individual differences

arising from differential experience with keyboards. Moreover, the software had to be contrived so as to flash a warning when a respondent accidentally pressed any other key. The "off-the-shelf" joysticks were sadly lacking in precision of construction such that the score of a respondent depended heavily upon which joystick (s)he was using.

We came up with a plan for a "response pedestal" which consisted of readily available electronic parts. The first design could probably have been constructed in a home workshop. A prototype of the device was obtained from a local engineer. (See Figure 1.) It had two joysticks, a horizontal and a vertical sliding adjuster, and a dial. The two joysticks allowed either left or right hand usage. The sliding adjusters permitted two-handed coordination tasks. The dial permitted respondent selections in a manner similar to the now popular "mouse" devices that are sold for personal computers.

The response pedestal had nine button-switches, each of which was to be used for particular purposes. Three buttons (BLUE, YELLOW, and WHITE) were located near the center of the pedestal and were used for registering up to 3-choice alternatives. Also near the center were two buttons (RED) which were mostly used to allow the respondent to step through frames of instructions and, for some tests, to "fire" a "weapon" represented in graphics on the screen.

Of notable interest was the placement of the button-switches which were called "HOME" with respect to the positions of other buttons used to register a differential response. The "HOME" buttons required the respondent's hands to be in the position of depressing all four of the "HOME" buttons prior to presentation of an item to which (s)he would respond. This, it is believed, offered advantages of control of attention and control of hand position for measurement of response latency. Using appropriately developed software, we were able to measure total response time but also to break it down into two parts: (1) "decision time" which is defined as the interval between onset of stimulus and release of the "HOME" keys, and (2) "movement" time which is the subsequent interval to the registering of a response. It was possible, where of interest, to tell fairly reliably whether the respondent used a left hand or a right hand to respond since (s)he almost invariably would release the "HOME" buttons on the side of the preferred hand first.

The rotary switch marked "SELECTOR" in Figure 1 was an inconvenience that was required by our initial choice of "game-adapter" for reading analog input. The game adapter initially chosen allowed only four inputs and the response pedestal had seven analog outputs: 2 inputs for each of two joysticks, two sliding adjusters, and one rotary adjuster called the "DIAL." The "SELECTOR" was used to select which analog devices were to be operative for a particular test item. The final design for the

Of notable interest was the placement of the button-switches which were called "HOME" with respect to the positions of other buttons used to register a differential response. The "HOME" buttons required the respondent's hands to be in the position of depressing all four of the "HOME" buttons prior to presentation of an item to which (s)he would respond. This, it is believed, offered advantages of control of attention and control of hand position for measurement of response latency. Using appropriately developed software, we were able to measure total response time but also to break it down into two parts: (1) "decision time" which is defined as the interval between onset of stimulus and release of the "HOME" keys, and (2) "movement" time which is the subsequent interval to the registering of a response. It was possible, where of interest, to tell fairly reliably whether the respondent used a left hand or a right hand to respond since (s)he almost invariably would release the "HOME" buttons on the side of the preferred hand first.

The rotary switch marked "SELECTOR" in Figure 1 was an inconvenience that was required by our initial choice of "game-adapter" for reading analog input. The game adapter initially chosen allowed only four inputs and the response pedestal had seven analog outputs: 2 inputs for each of two joysticks, two sliding adjusters, and one rotary adjuster called the "DIAL." The "SELECTOR" was used to select which analog devices were to be operative for a particular test item. The final design for the

response pedestal included a game-adapter with the capability of eight analog inputs and the "SELECTOR" switch was happily omitted.

Joysticks

Perhaps the greatest difficulty regarding the response pedestal design arose from the initial choice of joystick mechanisms. We soon discovered that joystick design is a complicated and, in this case, a somewhat controversial issue. Anyone who is contemplating the use of joysticks in a testing apparatus is well advised to not take the matter lightly. Variations in tension or movement can cause unacceptable differences in responding which defeat the goal of standardized testing..

While "high-fidelity" joystick devices are available, they can cost thousands of dollars apiece which was prohibitively expensive in the quantities that were to be required for this project. Additionally, we were not attempting to mimic any particular operational control device so there was no theoretical reason for high fidelity. The first joystick mechanism that was used in the response pedestals was an improvement over the initial "off-the-shelf" toys that predated the pedestals. It had no springs whatsoever so that spring tension would not be an issue. It had a small, light weight handle so that enthusiastic respondents could not gain sufficient leverage to break the

mechanism. It was inexpensive.

Unfortunately, this joystick had a "wimpy" feeling which was greatly lacking in "face-validity" (or, sometimes called "fist-validity") from the Army's point of view. It was felt that the joystick was so much like a toy that it would not command respect of the respondents. It was the contention of some of us that the device that we used had "construct fidelity" in that it would do a perfectly adequate job of testing the constructs that were targeted and that additional time and expense would be a waste. However, no amount of friendly persuasion would disuade the dissidents.

The joystick mechanism had to be changed. Joysticks of every conceivable variety and type of use were considered. We learned about viscous dampening, friction, tension, and even something called "stiction." Ultimately, a joystick device was fashioned with a light spring for centering and a sturdy handle with a bicycle handle-grip. It had sufficient "fist-validity" to be accepted by all (or almost all) and it was sufficiently precise in design that we were unable to detect any appreciable "machine" effects in fairly extensive testing.

## Software Development

We wish to turn attention now to the issues of software development. Obviously, there were no "package programs" available to administer computerized tests. The selection of strategy for organizing and programming the needed software was

to fall upon ourselves. We had three general, operational objectives in mind for the software to be produced: (1) as far as possible, it should be transportable to other microprocessors; (2) it should require as little intervention as possible from a test administrator in the process of presenting the tests to subjects and storing the data; and, (3) it should enhance the "standardization" of testing by adjusting for hardware differences across computers and response pedestals.

Primary Language

We chose to prepare the bulk of the software using the Pascal language as implemented by Microsoft, Inc. There were certain advantages to this in that Pascal is a common language and it is implemented using a compiler that permits modularized development and software libraries. As computer languages go, Pascal is relatively easy for others to read and it can be implemented on a variety of computers.

Some processes, mostly those which are specific to the hardware configuration, had to be written in IBM-PC assembly language. Examples of these include the interpretation of the response pedestal inputs, reading of the real-time-clock registers, calibrated timing loops, and specialized graphics and screen manipulation routines. For each of these identified functions, a Pascal-callable "primitive" routine with a unitary purpose was written in assembly language. Although the machine

specific code would be useless on a different type of machine, the functions were sufficiently simple and unitary in purpose so that they could be reproduced with relative ease.

Strategy

The overall strategy of the software development is worth considerable discussion. It quickly became clear that the direct programming of every item in every test by one person was not going to be very successful either in terms of time constraints nor in terms of quality of product. For the sake of making it possible for each researcher to contribute his/her judgment and effort to the project, it was necessary to plan so as to take the "programmer" out of the step between conception and product as much as possible.

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

blank during the construction of a complex display.

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

Thus, entire tests were constructed and experimentally manipulated by psychologists who could not program a computer.

The strategies for developing commands have evolved and improved over the period of development. Eventually, the commands became almost "language-like" with syntax forms analogous to some of the common statistical packages like SPSS or SAS that are available on "main-frame" computers.

Hardware Testing and Calibration

One of the most useful software developments relates to the testing and calibration of the hardware, necessary for purposes of standardization. A complete hardware testing and calibration process can be undertaken by test monitors each time a machine is powered up. It checks the timing devices and screen distortion, and calibrates the analog devices (joysticks, sliding adjusters, dial) so that measurement of movement will be the same across machines. It also permits the software adjustment of the height

to width ratio of the screen display so that circles do not become ovals or, more importantly, the relative speed of moving displays remains under control regardless of vertical or horizontal travel.

## Concluding Remarks

In the end, we were able to put together a portable, complete testing session lasting approximately 1-1/2 hours where very naive respondents can complete the test with little or no intervention from a test monitor. The data are automatically stored and "backed-up" on diskettes in a form readily transferrable to a "main-frame" for analysis. Except for occasional calibration or contingencies, the test monitor needs only to turn the computers on and put the respondents in front of them.

Finally, and perhaps most gratifying, we have found that the soldiers tested via this method have generally preferred computerized testing to paper-and-pencil testing. We have not gathered hard data on this aspect, but base our conclusions on observation of the soldiers while taking the battery and their comments to us after completing the battery. Perhaps this is due to novelty alone, but we feel it may also be due to the nature of the tests themselves plus the fact that the soldier, in large part, is in control of the testing process her/himself. The soldier controls the pacing of instructions for the tests and, for some tests, the pacing of item presentation. No

administrator tells her/him when to begin and when to stop, and (s)he is not in "lock step" with a larger group. We view this state of affairs as highly desirable for personnel selection testing.

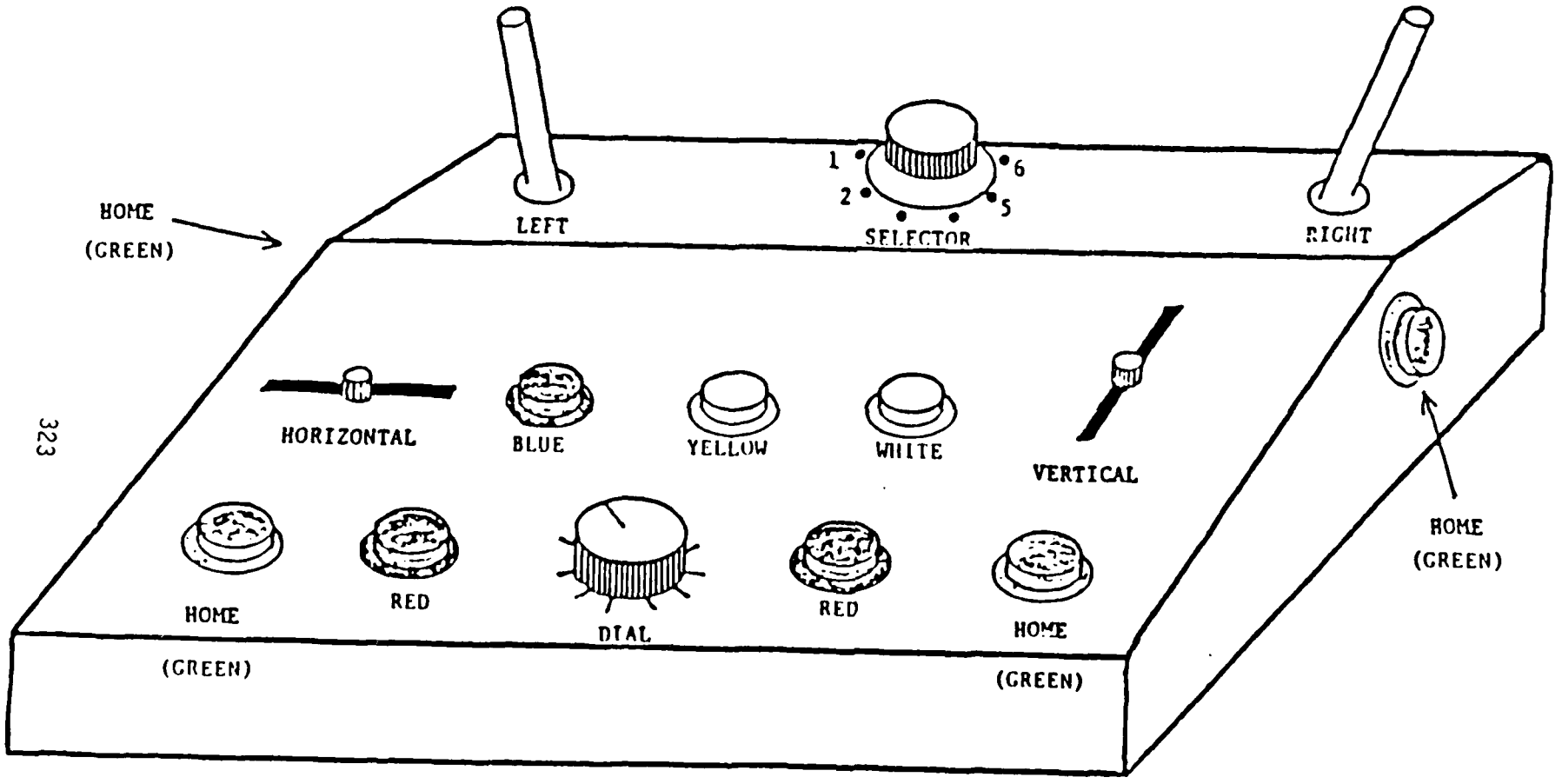



FIGURE 1. Custom-designed response pedestal

# MODELING THE SELECTION PROCESS TO ADJUST FOR RESTRICTION IN RANGE

Paul G. Rossmeissl
U.S. Army Research Institute

David A. Brandt
American Institutes for Research

August 1985

Presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine-year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 20263731A792 and is being conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the Army Research Institute, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

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.

# Modeling the Selection Process to Adjust for Restriction in Range

Paul G. Rossmeissl
U.S. Army Research Institute

David A. Brandt
American Institutes for Research

A central activity in validation research is the estimation of the regression of a criterion variable on one or more predictors in the applicant population. In most cases, however, the data for estimating this regression can be collected only for the subset of the applicants who are accepted for a job or training opening and have remained employed long enough to be tested on a criterion variable. This subset constitutes the *selectable* population. The selectable population is, in general, a biased sample of the population of applicants. Thus, sample statistics computed from such subsamples may be biased estimates of the parameters that are descriptive of the applicant population. To address this issue with the context of Project A research we considered a statistical adjustment that models the selection process in order to obtain unbiased estimates of regression parameters.

Treating this problem within the more general context of sampling bias, statisticians such as Heckman (1979), Goldberger (1972) and Muthén and Jöreskog (1983) have developed techniques to adjust for selection bias when it is possible to model the selection process.

The most common method of handling selection effects is to correct correlation coefficients for "restriction of range" using a Pearsonian correction. Texts by Lord and Novick (1968) and Allen and Yen (1977) describe these methods and the assumptions behind them.

The statistical methods that we evaluate in this paper differ from the psychometric methods in one crucial respect. The psychometric methods assume that the regression of the criterion on the predictor is the same in the applicant and the selectable populations. On the basis of this assumption, they adjust the correlation coefficients using the observed variances in the two populations. That is, the correction is for "restriction in range," and nothing more. In fact, the sample regression obtained from the selectable population could also be regarded as the "corrected" regression. This is true in the sense that the correction for correlation coefficients assumes that the regression in the selectable population is the same as the regression in the applicant population.

On the other hand, the statistical methods relax the assumption of equal regressions in the applicant and selectable populations. These models assume that selection affects both the correlation and the regression coefficients. This is an important difference between the two approaches, because it is likely that the

regression in the selectable population has a higher intercept and a flatter slope than the regression in the applicant populaton. Therefore a method that is capable of making that adjustment is of considerable value in validation research.

The basis of the adjustment for selection effects is the assumption that the true selection variable is linearly related to one or more observed variables. The true selection variable, denoted by $\eta$, is not observed, but the variables that are related to selection are observed. It is assumed that the researcher wishes to estimate the linear regression equation of a dependent variable, y, on a vector of x's but y is observed only if $\eta$ is greater than a threshold value, $\tau$.

In order to model the selection process, data from both successful and unsuccessful applicants are concatenated and a binary variable is generated that indicates whether an applicant was selected. Then the probability that a given applicant was selected is estimated statistically. Heckman (1979) used probit analysis for this purpose. In the probit analysis, all the variables thought to be related to selection are included as predictors. A function of the estimated probability of selection is then included as a predictor in the regression equation of interest. This augmented equation then is estimated by a technique such as ordinary or generalized least squares. If the distributional assumptions can be satisfied, unbiased estimates of the regression equation of interest will be obtained. Other estimation methods such as ridge or Bayesian regression can also be used.

Muthén and Jöreskog (1983) proposed a maximum likelihood estimator of the system of equations. While the Muthén-Jöreskog estimation method is more attractive from a theoretical point of view, practical problems have been encountered in its application. The numerical optimization may be nontrivial because the shape of the likelihood surface is frequently quite complicated, with many local minima. Also, Muthén's computer program is not available for general distribution. The program that implements the Heckman estimation procedure, however, is commercially available.

The purpose of this paper is to address the feasibility of using the Heckman procedure for addressing the problem of selection bias in connection with Project A. Analyses of artificial data presented by Muthén and Jöreskog indicate that the Heckman adjustment performs well under the conditions that it was designed to address. However, test validation research often falls short of this idea in several respects. In particular, small sample size and multicollinearity may be problems. The smallest sample size investigated by Muthén and Jöreskog was 1000. Validation sample sizes are often as small as 100. There is additional concern about multicollinearity in connection with the Heckman model. We will use artificial data to compare adjusted regression estimates obtained using Heckman's procedure to unadjusted coefficients obtained using OLS. We will also investigate the impact of different selection ratios on the parameter estimates.

In theory, Heckman's procedure "works" when the selection process can be modeled adequately. That is, in simple selection situations in which the selection variables are measured, the Heckman procedure should properly adjust for

selection bias. It is less clear what the procedure will do when some of the variables that contribute to selection are unmeasured.

Estimation of parameters from small samples poses additional problems. First, the standard errors of the adjusted regression parameters are, in general, larger than the standard errors of unadjusted regression estimates. This reflects the fact that the selection variable is not observed but is estimated from the same cases on which the regression is obtained.

Collinearity is also an important concern. Heckman's procedure attempts to correct for specification error by adding a term to the regression that "compensates" for nonrandom selection. This statistic, lambda ($\lambda$), is derived from a probit regression of selection (0 or 1) on the variables that are believed to be relevant to selection. Ordinarily, this set of variables includes all the variables in the regression that is to be adjusted. This means that $\lambda$ is, to some extent, a non-linear function of the same variables that are already in the regression equation. If selection is only a function of the variables in the OLS equation, collinearity can be an extremely serious problem when estimation is based on small samples.

## Method

### The Heckman model

This method assumes that the researcher is interested in the regression

$$y = X_1 \beta_1 + \epsilon_1$$

and that selection is a function of the unobserved variable, $\eta$,

$$\eta = X_2 \beta_2 + \epsilon_2$$

That is, the dependent measure, $y$, is only available for cases in which $\eta > \tau$. The task is to adjust the coefficients, $\beta_1$, for the effects of nonrandom selection.

Heckman uses a two-step estimation procedure. First, a probit regression is used to model the selection process. The researcher performs a probit analysis of selection (i.e., selected or unselected) on all the variables believed to contribute to selection. These variables make up the $X_2$ matrix above. From this analysis, the probability that each individual is selected is obtained. A function of this probability, $\lambda$, is entered into the regression equation of interest as an additional predictor[1]. In theory, this predictor adjusts for the bias caused by nonrandom selection.

The outcome of the analysis is a table of regression coefficients adjusted for nonrandom selection together with the covariance matrix of these estimates. Heck-

[1]Lambda is the inverse of Mill's ratio. It is a monotone decreasing function of the probability that an observation is selected into the sample.

man (1979) worked out analytic expressions for these standard errors. In general, they can be expected to be larger then the standard errors obtained from normal theory that assumes random selection from an indefinitely large population.

These calculations have been implemented in the computer program, LAMBDA. This program is available from Scientific Software, Inc. (formerly International Educational Services). The work reported here was done using this program.

This simulation was designed to examine the performance of the Heckman correction in simple cases. We studied the regression of an outcome variable on a single predictor. We varied the following factors:

- Sample size (100, 300, 600);
- Selection ratio (50% eligible, 33% eligible, 16% eligible);
- The compositon of the selection variable;
- The degree to which the variables contributing to selection were included in the probit analysis.

Simulation specifications

All data were generated using the multivariate normal generator, GGNSM, in the International Mathematical and Statistical Libraries (IMSL) subroutine package. For reasons of simplicity, we restricted our attention to the regression of an outcome variable on a single predictor. The behavior of the Heckman procedure in multiple regression situations will be studied in future work. The simple linear regression equation case, however, is relevant to much validation research. It is the method used to evaluate the predictability of operational and proposed predictor tests, and to investigate predictive bias of of such tests.

We generated a predictor variable, X and two possible selection variables, $S_1$ and $S_2$. $S_1$ is correlated .5 with the predictor, X, and $S_2$ is correlated .25. They represent other factors related to selection. Two disturbance terms, $\epsilon_1$ and $\epsilon_2$, that are uncorrelated with X, $S_1$, and $S_2$, were also generated. We created the dependent variable from the equation

$$y = X + \epsilon_1$$

and studied this model for two possible selection variables:

$$\eta^{(1)} = X + S_1 + \epsilon_2, \text{ and}$$

$$\eta^{(2)} = X + S_1 + S_2 + \epsilon_2.$$

The first selection variable is perhaps the simplest case that is at all realistic. It assumes that selection is on the basis of a predictor test plus another

variable that is moderately correlated with the predictor. The second selection variable assumes that selection is on the basis of the predictor plus two variables that are correlated with it.

Finally, it is necessary to make assumptions about the availability of $S_1$ and $S_2$ to the researchers. We ran simulations that assumed that:

- Selection is on X and $S_1$ but the researchers only have X;
- Selection is on X and $S_1$ and the researchers have both X and $S_1$;
- Selection is on X, $S_1$, and $S_2$ but the researchers only have X and $S_1$.

The first condition simulates the case in which the researchers only have the predictor test available, but selection is also on the basis of an unmeasured variable correlated with it. The second condition is included as a baseline against which the results from the first conditions can be compared. Finally, the last condition represents the case in which some of the selection process is measured but an unmeasured variable with a low correlation with the predictor test also contributes to selection. Comparisons of the second condition with the first and last will indicate the effects of failing to model the selection process properly.

This research is not a Monte Carlo simulation. For each condition, we generated an appropriate artificial dataset and performed the unadjusted and adjusted regressions. Our intent was to get an initial sense of the behavior of the adjustment procedure for the cases of interest to us. Subsequent work based on these initial results will use more rigorous procedures.

## Results

Because this research does not generate a distribution of $\hat{\beta}$'s for each condition, some of the variability in our estimates is due to sampling error. Therefore, we can only be confident about sizeable and consistent effects. This limitation is satisfactory for us at this stage of our work. The reader is cautioned that more subtle trends in our data should be investigated with formal Monte Carlo simulations.

Table 1 contains estimated regression coefficients ($\hat{\beta}$) and their standard errors obtained from unadjusted and adjusted regressions. In all cases, the population $\beta$=1.0. Different selection ratios were obtained by varying the threshold that determines selection. Fifty per cent selection corresponds to setting the threshold at zero (i.e., $\tau$=0); a 33% selection rate is obtained by setting $\tau$=1, and a 16% selection rate is obtained by setting $\tau$=2. The two sets of columns on the right report on adjusted estimates for the case in which the selection variable is $\eta^{(1)}$. The first set of adjusted estimates was obtained by using only X in the probit analysis; the second set uses both X and $S_1$. This table permits us to compare the

effects of sample size, selection ratio, and the degree to which the researcher has measured the variables relevant to selection.

The following features are worth noting. First, selection affects the uncorrected estimates in the expected way. In all cases, the uncorrected coefficients underestimate $\beta$, and the effect is more pronounced for the higher selection ratios. When $\tau$ is 2 or 3 and the sample size is 300 or more, the unadjusted estimates of $\beta$ are significantly lower than 1.0. This is a clear indication that some form of adjustment is needed.

The most dramatic effect in Table 1 is due to the number of variables included in the probit analysis. When a $S_1$ is not available to the researcher, the Heckman procedure does not perform well. The first set of adjusted estimates were obtained using only X in the probit analysis. The standard errors of the estimates are roughly 8 times the standard errors of the unadjusted estimates. When the sample size is 100, the degree of inflation is even worse. The only case that could be considered at all reasonable is for a high selection ratio and high N (i.e., 600). In this instance, $\hat{\beta}$=1.13, but its standard error is almost nine times the standard error of the corresponding unadjusted estimate.

The set of estimates on the right of Table 1 illustrates the importance of measuring all the variables relevant to selection. These estimates were obtained under by specifying the correct set of variables in the probit analysis. For this case, Heckman consistently overcorrects, but in nearly all cases the adjusted $\hat{\beta}$'s are not significantly different from 1.0. More importantly, the standard errors are only about $1\frac{1}{2}$ times the standard errors of the unadjusted estimates.

Even when all the variables relevant to selection are known, a sample size of 100 may not be adequate. For the highest selection ratio studied here, the adjusted $\hat{\beta}$ based on 100 observations is significantly greater than 1.0. A Monte Carlo simulation would probably be required to determine whether Heckman estimates could be relied upon in this case.

If we disregard the results for N=100, it appears that the Heckman estimator does better as the selection ratio increases. This is "good news" in the sense that it is precisely this case in which a good adjustment procedure is most needed. The last two rows of Table 1 illustrates the case in which Heckman does best *and* is of greatest practical value. In these rows the unadjusted $\hat{\beta}$'s are significantly attenuated but the adjusted estimates are excellent. Of course, it is crucial that the researcher measure the variables relevant to selection.

We also generated a second selection variable, $\eta^{(2)}$, to investigate further the effects of misspecifying the probit model. This selection variable is a function of X, $S_1$, and $S_2$; we are interested in the case in which only one of the additional variables related to selection is measured. Table 2 contains the unadjusted and adjusted estimates when the selection variable is $\eta^{(2)}$ but only X and $S_1$ are available to the researcher.

From Table 2 it is clear that not knowing $S_2$ does not have the devastating effects on precision that were seen in Table 1 when $\eta^{(1)}$ was not properly modeled. Standard errors for adjusted $\hat{\beta}$'s are only about $1\frac{1}{2}$ times the standard errors for the corresponding unadjusted $\hat{\beta}$'s. However, the correction is not quite as good as when $\eta^{(1)}$ was properly modeled. Heckman's procedure seems to overcorrect whenever the unadjusted $\hat{\beta}$ is larger than about .88. Also, the performance of the adjustment for extreme selection ($\eta > 2$) is not better than the other cases. These results suggest that the Heckman procedure gives reasonable results for some cases in which all variables relevant to selection are not measured. This is encouraging to Project A because it is not realistic for us to measure all variables that are relevant to the Army's selection and classificaton process.

## Summary

Our findings suggest several things. First, the Heckman procedure does not perform well when only variables in the regression of interest are included in the probit analysis. In this case, the standard errors of the adjusted estimates are unacceptable. However, we do find that if some but not all of the variables related to selection can be added to the probit analysis, the adjusted estimates are reasonable.

As we expected, a sample size of 100 is probably not large enough to be relied upon. For Project A, this means that we may not be able to use the Heckman procedure to investigate predictive bias in many cases because the subgroup N's are of that size.

The results using $\eta^{(2)}$ were encouraging. They indicate that the adjustment may perform reasonably even when some variables related to selection are unmeasured. However, it is crucial that some additional variables related to selection that do not appear in $\mathbf{X}_1$ be used in the probit analysis. We found that the standard errors of parameter estimates were acceptably low and the adjusted $\hat{\beta}$'s may be an improvement over the unadjusted estimates. A more rigorous simulation and some analyses of real data are needed to determine whether the data being collected by Project A are suitable for this procedure.

# References


Allen, M. J., & Yen, W. M. (1979) *Introduction to measurement theory.* Belmont, CA: Brooks-Cole.

Goldberger, A. S. (1980) *Methods for eliminating selection bias.* Department of Economics, University of Wisconson, Madison.

Heckman, J. J. (1979) Sample selection bias as a specification error. *Econometrica, 47(1),* 153-161.

Lord, F. M., & Novick, M. R. (1968) *Statistical theories of mental test scores.* Menlo Park, CA: Addison-Wesley.

Muthén, B., & Jöreskog, K. G. (1983) Selectivity problems in quasi-experimental studies. *Evaluation Review, 7(2),* 139-174.

Table 1

Comparison of Uncorrected and Corrected Regression Coefficients for $\alpha = 0$; $\beta = 1$

Selection variable: $\eta = x + S_1$

| | Uncorrected | | Corrected | | | |
|---|---|---|---|---|---|---|
| Selection rule: $\eta > 0$ | | | Probit analysis excludes $S_1$ | | Probit analysis includes $S_1$ | |
| N | $\hat{\beta}$ | S.E. | $\hat{\beta}$ | S.E. | $\hat{\beta}$ | S.E. |
| 100 | .94 | .11 | -.52 | 7.16 | 1.30 | .19 |
| 300 | .92 | .06 | .96 | .54 | 1.30 | .10 |
| 600 | .79 | .04 | .81 | .29 | .90 | .06 |
| Selection rule: $\eta > 1$ | | | | | | |
| 100 | .61 | .11 | 2.49 | 2.04 | .98 | .17 |
| 300 | .79 | .07 | -.08 | .84 | 1.11 | .11 |
| 600 | .85 | .05 | 1.51 | .39 | 1.09 | .08 |
| Selection rule: $\eta > 2$ | | | | | | |
| 100 | .79 | .12 | 1.56 | 1.92 | 1.42 | .17 |
| 300 | .74 | .06 | 1.32 | .75 | 1.07 | .09 |
| 600 | .71 | .04 | 1.13 | .35 | 1.02 | .06 |

Table 2

Comparison of Uncorrected and Corrected Regression Coefficients for $\alpha = 0$; $\beta = 1$

Selection variable: $\eta = x + S_1 + S_2$

| Selection rule: $\eta > 0$ | Uncorrected | | Corrected[1] | |
|---|---|---|---|---|
| N | $\hat{\beta}$ | S.E. | $\hat{\beta}$ | S.E. |
| 100 | .94 | .11 | 1.14 | .15 |
| 300 | .79 | .06 | .92 | .10 |
| 600 | .96 | .04 | 1.13 | .06 |
| Selection rule: $\eta > 1$ | Uncorrected | | Corrected[1] | |
| 100 | .76 | .11 | .83 | .16 |
| 300 | .99 | .06 | 1.06 | .09 |
| 600 | .82 | .04 | .92 | .06 |
| Selection rule: $\eta > 2$ | Uncorrected | | Corrected[1] | |
| 100 | .75 | .12 | .77 | .15 |
| 300 | .81 | .05 | .98 | .08 |
| 600 | .88 | .04 | 1.12 | .05 |

[1] The variables x and $S_1$ are included in the probit analysis.

# COMPARING WORK SAMPLE AND JOB KNOWLEDGE MEASURES

by

Michael G. Rumsey
U.S. Army Research Institute

William C. Osborn and Patrick Ford
Human Resources Research Organization

August 1985

Presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

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

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

# Comparing Work Sample and Job Knowledge Measures

Michael G. Rumsey

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

William C. Osborn and Patrick Ford

Human Resources Research Organization

In our Army Selection and Classification project, commonly referred to as Project A, we are concerned with developing as comprehensive a performance measurement system as possible. To that end, we have developed three different kinds of measures--ratings, work sample measures and job knowledge measures. The role of ratings will be discussed in the following paper. Here we focus on the two remaining testing methods--work sample and job knowledge tests.

As a measure of job proficiency, work sample tests enjoy a level of acceptance second to none. It has been suggested that, short of measurement in an actual job situation, a work sample test has the highest fidelity of any type of measure (Vineberg & Taylor, 1976). Ideally, a work sample test represents the actual steps performed in a job and provides an objective measure of whether or not such steps are performed correctly. If this ideal is met, then an individual's score on the work sample test is an unparalleled measure of how well the individual can successfully satisfy the requirements of a particular job.

Despite their extraordinarily high level of credibility, work sample measures are very infrequently used as a component of a performance appraisal system. The most common explanation for this omission is the enormous expense associated with such testing. To properly conduct work

sample testing, one must provide the equipment that is used on the job, at least one observer to determine if the job is being performed correctly, and make available sufficient time to insure that all essential components of the job are executed. In most cases, the price of such testing is simply more than the user feels can be afforded.

Given the perception that work sample measures represent the highest standard in proficiency measurement, and given the often prohibitive expense of work sample measurement, it follows that all other measures would be assessed in terms of the extent to which they could adequately substitute for work sample measures.

Indeed, this represents the prevailing manner in which knowledge tests are viewed. Foley (1977) reviewed findings from several studies employing job knowledge and job performance tests in maintenance jobs and concluded that the correlations between such measures were not sufficiently high to warrant substituting knowledge tests for work samples. Vineberg and Taylor (1972b) compared job knowledge tests and job sample tests in four Army occupations and found sufficiently high correlations to justify their conclusion that (p. 19) "job knowledge tests can be appropriately substituted for job sample tests, when a job contains little or no skill components and when only knowledge required on the job is used in the test."

Despite the apparent reasonableness and seductive simplicity of the substitutability approach, it rests on assumptions that cannot be proven and may not always be justified. It is difficult to argue with the proposition that work sample measures represent the most appropriate standard when the job can be easily represented, tested and scored in a work sample

mode. Vineberg and Taylor (1972b, p. 17) suggested that: "Where job performance relies almost solely on a skill, job sample tests, or some other variety of performance measure, are essential." Yet these authors noted that, where knowledge is an important part of the job, both knowledge tests and performance tests are appropriate. To use their example, a knowledge test is better suited to assess an automobile driver's knowledge of driving rules and road signs than is a performance test. Unless there is clear evidence that a work sample test can adequately cover all aspects of job performance, one cannot rule out in advance the possibility that a knowledge test may provide a unique, valid contribution to an overall assessment of an incumbent's job proficiency.

Such a perspective has clear implications for how we look at relationships between scores on work sample and knowledge tests. The correlations we obtain do not tell us how well the knowledge test score approximates the "true score" recorded on the work sample test; neither type of test can be presumed in advance to provide a perfectly accurate representation of truth. However, we can examine the correlations in the context of what we know about the jobs being tested and what we know about the strengths and weaknesses of the tests themselves to determine whether our expectations about the degree of interchangeability of each type of measure are confirmed and, if not, what kinds of revisions might be needed, either to our measures so that they might better fit our expectations, or to our expectations themselves. The information might not be so neatly interpretable as it would be if we were using the substitutability approach, but it is no less valuable.

A relatively small number of investigations have examined the rela-

tionship between work sample and knowledge measures. Several of these were noted in an article by Hunter (1983), who was examining the contribution of job knowledge, work samples and ability to performance ratings. Hunter found an average correlation of .67, corrected for attenuation, between work sample and knowledge measures. Without correction for attenuation, the average correlation was .52. This review contrasts dramatically with that conducted by Foley (1974), who found uncorrected correlations ranging from .10 to .55 in maintenance jobs.

The evidence indicates that there is a substantial relationship between work sample and knowledge measures, although the relationship is an imperfect one. The evidence also shows substantial variability in the results obtained. This variability might be attributed to a variety of factors, including statistical artifacts. One purpose of the present investigation was to examine the extent to which type of work sample measure and type of occupation account for the variability observed.

Asher and Sciarrino (1974) identified two types of work sample tests: Motor and verbal. These were defined as follows (p. 519): "A work sample was identified as "motor" if the task was a physical manipulation of things, as for example, tracing a complex electrical circuit, operating a sewing machine, making a tooth from plaster, or repairing a gear box. A work sample test was classified as "verbal" if there was a problem situation that was primarily language-oriented or people-oriented."

If we refine Asher and Sciarrino's definition slightly and focus only on those verbal work samples that are primarily language-oriented, it is readily apparent how their distinction between the two types of work samples would be important when considering correlations between job knowl-

edge and work sample tests. A motor work sample test differs from a job knowledge test in two ways: Only the job knowledge test is verbal and only the work sample requires actual task performance. Only the latter difference applies when the work sample is itself verbal. Thus, correlations with written tests should be higher for verbal than for motor work samples.

Vineberg and Taylor (1972b) have suggested that the distinction between knowledge and skill is critical with respect to expectations regarding correlations between knowledge and work sample measures. They noted that skill, unlike knowledge, can only be acquired through practice, and identified four major skill categories: Perceptual, motor, cognitive and social. Job knowledge tests are presumably best suited to measure job knowledge; work sample 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 jobs in which skill is an important requirement, the correspondence should be lower.

A direct test of Vineberg and Taylor's (1972b) thesis would be difficult, requiring knowledge about the specific skill requirements of each job. Some preliminary work is being done in Project A to examine the relationship between the judged skill requirements for a particular task and the relationship between work sample and job knowledge tests for that task, but the relevant data analyses for that effort are not yet complete. In the meantime, a grouping of jobs based on underlying cognitive requirements, which may provide some indication of the skills needed for these jobs, is available.

The Armed Services Vocational Aptitude Battery (ASVAB) is composed of a set of cognitive tests used in the selection and placement of applicants for military service. An Aptitude Area composite defines a group of such tests which optimally predict training performance in a specified set of Army occupational specialties. McLaughlin, Rossmeissl, Wise, Brandt and Wang (1984) recently determined that four composites were essentially sufficient for grouping jobs on the basis of aptitudes measured by the ASVAB. These composites were labelled: Clerical, skilled technical, operations and combat. They served as a basis for the present exploratory investigation of the effect of type of occupation on relationships between work sample and knowledge tests.

Past findings have been systematically summarized to determine the appropriate context in which Project A findings should be evaluated. Meta-analytic procedures described by Hunter, Schmidt and Jackson (1982) have been used to examine overall relationships between work sample and knowledge tests and relationships within specific occupational groupings. The Army's Project A is unique in terms of the variety of occupations for which work sample and knowledge tests have been systematically developed. Because the same test development methods have been applied consistently across occupations, this project offered an unprecedented opportunity to consider the replicability of relationships suggested by earlier investigations. Major discrepancies between Project A and earlier findings provided a basis for examining both Project A procedures and earlier investigations to identify possible explanations for such discrepancies.

In order to compare the findings from this project to findings from previous efforts, we needed some procedure for cumulating both sets of

findings. Hunter, Schmidt and Jackson (1982) have cautioned against drawing inferences concerning apparent discrepancies in findings drawn from accumulated studies without subjecting such findings to prescribed meta-analytic techniques designed to identify and control for statistical artifacts which might account for such discrepancies. We have in this effort used such procedures to examine whether there is a basis for identifying a moderating variable which might explain the apparently discrepant correlational findings previously observed. We have used similar procedures to compare correlational findings across different occupations in Project A. Finally, we have used these methods to compare findings between Project A and previous efforts and, as appropriate, cumulate findings across both.

## Method

### Literature Review

Identifying Sources. An intensive search was conducted to identify all investigations which provided comparisons between work sample and knowledge measures. Certain criteria were applied before an investigation could be included in this review. First, the availability of a Pearsonian correlation coefficient or equivalent measure for the between-methods comparison was required. This criterion resulted in the exclusion of one investigation (Grings, 1953), which reported a rank order coefficient. Second, investigations which intentionally manipulated score variability by special training procedures were eliminated. This criterion also resulted in the exclusion of one investigation (Osborn & Ford, 1977). A third source (Saupe, 1955) was not obtained in time to be included in this review. Since the sum of the sample sizes for the three excluded inves-

tigations was 113, the impact of these exclusions was minimal.

The literature review identified 14 investigations reporting 19 separate comparisons, with a total sample size of 4151. These findings are shown in Table 1. Ten of the 19 comparisons shown were summarized by Hunter (1983). Some investigations used multiple measures of the same type (job knowledge or work sample) and some used multiple samples. Correlational and reliability values shown represent mean values when multiple values were reported.

All reliability estimates reported for work sample or job knowledge tests were of the internal consistency type. Each was derived using one of the following types of computational techniques: Coefficient alpha, Kuder Richardson formula 20 (KR-20) or split half.

Correlations with months on the job partialled out were used when available, otherwise zero-order correlations were used. For only four comparisons, those reported by Vineberg and Taylor (1972a, b), were both zero-order and partial correlations available. In those cases (as noted by J. E. Hunter, personal communication, July 10, 1985), the extreme variability in subject time on the job (from one month to over 20 years) relative to mean time spent on the job (the majority of examinees had been on the job less than 15 months) presented a unique circumstance which dictated the use of the partial correlation to maximize comparability with other investigations.

Cumulating Findings. Hunter, Schmidt and Jackson (1982) have recommended a specific set of procedures for combining correlational findings across studies. These procedures will be summarized here, with comments regarding the manner in which these procedures were applied in this

Table 1

Work Sample and Knowledge Tests: Previous Results

| Authors | Occupation | Sample Size | Reliability | | Correlations | |
|---|---|---|---|---|---|---|
| | | | $r_{kk}$ | $r_{ww}$ | $r_{kw1}$ | $r_{kw2}$ |
| Schoon (1979) | Medical Laboratory Worker | 160 | 91 | 95 | 72 | 77 |
| Vineberg and Taylor (1972b) | Supply Specialist | 380 | 92 | — | 65 | 80 |
| van Rijn and Payne (1980) | Firefighter | 210 | 78 | 77 | 62 | 80 |
| Campbell et al. (1973) | Cartographer | 443 | 88 | 49 | 52 | 79 |
| Vineberg and Taylor (1972b) | Cook | 366 | 84 | -- | 50 | 65 |
| Vineberg and Taylor (1972b) | Armor Crewman | 368 | 81 | — | 49 | 65 |
| Vineberg and Taylor (1972b) | Armor Repairman | 360 | 76 | -- | 49 | 67 |
| Corts et al. (1977) | Customs Inspector | 186 | 67 | 80 | 49 | 67 |
| Maier (1982) | Infantry Rifleman | 193 | -- | -- | 44 | 57 |
| Livingston (1978) | Radiologic Technologist | 140 | — | -- | 42 | 55 |
| O'Leary and Trattner (1977) | Tax Investigator | 292 | 64 | 78 | 41 | 58 |
| Trattner, et al. (1977) | Claims Examiner | 233 | 81 | 72 | 34 | 45 |
| Evans and Smith (1953) | Electronics Technician | 57 | 81 | 44 | 34 | 57 |
| Williams and Whitmore (1959) | Electronics Maintenance | 189 | 92 | 88 | 33 | 37 |
| Crowder, et al. (1954) | Radar Mechanic | 119 | -- | 69 | 33 | 42 |
| Engel and Kehder (1970) | Vehicle Mechanic | 30 | 91 | 82 | 27 | 31 |
| Maier (1982) | Automotive Mechanic | 131 | -- | -- | 26 | 34 |
| Brown, et al. (1959) | Field Radio Repair | 235 | 83 | 70 | 20 | 26 |
| Maier (1982) | Ground Radio Repair | 59 | -- | -- | 12 | 16 |

Note. Table partially adapted from tables presented in "A Causal Analysis of Cognitive Ability, Job Knowledge, Job Performance, and Supervisor Ratings," by J. E. Hunter, 1983. In T. Landy, S. Zedeck and J. Cleveland (eds), Performance Measurement and Theory, Hillsdale, NJ: Lawrence Erlbaum, pp. 260-261. Reliabilities shown for the Campbell, et al. (1973) investigation were based exclusively on this source; all other data have either been independently verified or obtained from other sources.

Explanation of terms:

$r_{kk}$ - Reliability of job knowledge test

$r_{ww}$ - Reliability of work sample

$r_{kw1}$ - Unadjusted correlation between knowledge test and work sample

$r_{kw2}$ - Correlation adjusted for attenuation

review.

The first step involves correcting three sources of error from individual investigations: Sampling error, error of measurement, and range restriction. The correction for sampling error was applied in this review and raised no apparent problem. The correction for error of measurement deserves some comment. Three types of reliability indices are available for measuring error of measurement: Interrater, test-retest, and internal consistency. None of these is entirely adequate with respect to work sample proficiency testing. Interrater reliability addresses only error associated with scorer behavior. Test-retest reliability is easily contaminated by memory associated with earlier responses and learning occurring between testing sessions. Internal consistency reliability is most appropriate when the objective is to measure factorially pure traits and items are mutually independent. The factorial purity of work sample proficiency tests will vary according to the content of the job and item independence will be violated when an examinee's failure to perform the initial steps of a task make it impossible for that examinee to perform the final steps as well.

Internal consistency indices are those typically reported in the literature for both work sample and job knowledge tests. This review will examine correlations corrected using such indices, recognizing that uncorrected coefficients are underestimates of the true relationship, but uncorrected coefficients will be carefully examined as well because the corrected coefficients are likely themselves biased by an error component that cannot be effectively quantified.

A systematic correction for range restriction has not been attempted in this review. Such corrections are not typically performed when the proficiency measures are both designed for and administered to a job incumbent population. This does not mean that such a correction would be necessarily inappropriate for obtaining the information sought here. Variations in selection standards across investigations may well have impacted upon the range of ability in the samples tested in such a way as to contaminate the correlational findings. However, the information needed to systematically examine the impact of range restriction upon such findings was not available in the investigations reviewed.

The adjustments to correlation coefficients to compensate for error of measurement can be accomplished in one of two ways. If each correlation can be individually adjusted, then that is the recommended procedure. If there is missing data with respect to reliability or range restriction in any of the reported investigations, then Hunter et al. (1982, pp. 74-87) recommend using information on the distribution of these variables in formulas which they have provided to correct the overall mean and standard deviation of the correlations.

The use of distributional artifact data rather than artifact data from individual investigations does have one limitation. If error of measurement is not randomly distributed but is instead positively or negatively associated with the magnitude of uncorrected correlations, use of distributional data may produce a substantial underestimate or overestimate of the variance of the corrected correlations. For this reason and to enhance the comparability of computations made on literature review data, where some artifact values were missing, with computations

made on Project A data, where no artifact values were missing, a modification to the Hunter et al. (1982) approach was used here. For each missing value, the mean value reported for that variable in investigations where data was available was inserted. Each correlation was then individually corrected using either available artifact data or the substitute mean value. Mean reliability values used for individual corrections were .83 for job knowledge tests and .71 for work sample tests.

Once all reasonable steps have been taken to remove error, an average correlation coefficient can be computed. Hunter, et al. (1982) suggest that the variance across studies be analyzed before any search for moderator variables is initiated. Only if the variance, appropriately corrected, is positive and substantial is a search for moderator variables warranted. If moderator variables are suspected, then correlations can be grouped accordingly and corrected variances and correlations examined within the new groupings.

Work Sample Categories. Using a modified version of Asher and Sciarrino's (1974) distinction, work samples in each investigation reviewed were classified as motor or verbal based on the description of the work sample provided. If a work sample consisted of both motor and verbal elements, the classification was based on which type of element was predominant.

Occupational Categories. The Army occupations that corresponded to the clerical, skilled technical, operations and combat groupings were identified in the report by McLaughlin, Rossmeissl, Wise, Brandt and Wang (1982). The Occupational Conversion Manual (Department of Defense, 1982) was used to identify Army equivalents to non-Army jobs, then the non-Army

jobs were similarly grouped.

In the combat grouping, because of the relatively large number of investigations available, examination of a further distinction was possible. The combat cluster identified in the McLaughlin, et al. (1982) report consolidated four clusters identified in earlier research: Combat, field artillery, general maintenance and electronics repair. Some early clustering work by R.G. Hoffman which differentiates jobs on the basis of judgments of similarities of performance requirements (personal communication, April 18, 1985) provided some indication that the distinction between combat and field artillery may not be a clean one; accordingly, the two combat arms sub-clusters were combined here for comparative purposes. The Hoffman work did maintain electronics repair as a separate cluster; accordingly, the combat cluster was subdivided into combat arms and electronics repair categories for purposes of the present investigation. Since only one (van Rijn & Payne, 1980) of the investigations reviewed examined an occupation falling within the general maintenance sub-category, that classification was ignored in this review. As with the other classification assignments, the McLaughlin (1982) data and the Occupational Conversion Manual (Department of Defense, 1982) were used to determine the appropriate sub-category for each job in the combat cluster.

Project A

Occupations Covered. Work sample and job knowledge measures were developed for nine military occupational specialties. These specialties were selected to be as representative of the full set of Army MOS as possible, to represent important Army MOS, and to be sufficiently large to allow longitudinal comparisons of predictor and performance measures. To

insure representativeness, various means of grouping MOS were considered. These groupings consisted of the Army's administrative divisions, known as Career Management Fields (CMF), a grouping based on cognitive predictors, known as Aptitude Areas, and a grouping based on judgments of similarity of job conduct, which is described in Rosse, Borman, Campbell, and Osborn (1983). The MOS were chosen to cover as many components of each as possible. The MOS selected were infantryman (11B), cannon crewman (13B), tank crewman (19E), radio teletype operator (31C), vehicle mechanic (63B), motor transport operator (64C), administrative specialist (71L), medical specialist (91A) and military police (95B).

Tasks Covered. Both work sample and knowledge tests were designed to cover discrete components of Army jobs known as tasks. Experience suggested that 30 tasks per job would provide reasonable job coverage. It was determined that, for field test purposes, 15 tasks would be tested in a work sample mode and all 30 would be tested in a knowledge mode. That would equate to approximately four hours testing time per individual soldier for both the work sample and knowledge tests.

Since a given job may have as many as several hundred tasks, some procedure was needed to identify that set of 30 tasks which would best represent the overall domain. The first step needed was to identify those tasks which constituted the domain. This involved consulting Army sources of task information, reconciling these sources with one another, and verifying the tentative task domain thus derived with Army subject matter experts for a particular MOS.

The next step required was to reduce the total task domain to a list of 30. This reduction was based on a specified set of criteria. First,

the tasks needed to broadly cover the content of the job. Second, the tasks should be the relatively more important ones. Third, the tasks should permit some variability of performance. Systematic judgments on each of these criteria were obtained from Army subject matter experts. Frequency of task performance was also considered; such information was obtained from official Army sources. Information on task clusters based on similarity of performance requirements, task importance, task performance variability and task frequency was then considered by project staff and, in some cases, by Army subject matter experts, who, using a modified Delphi procedure, reached consensus on the 30 tasks to be tested per MOS. Where Army subject matter experts were not involved in the task selection process, the outcome of the process was reviewed by representatives from the Army proponent agency, who provided concurrence with the tasks selected.

Test Development. Fifteen of the 30 tasks were selected for work sample testing based on such factors as number of cued steps and degree of skill required. Available Army reference materials were used to determine steps involved in performing each task. Work sample tests were developed to score the soldier on whether each such step was correctly performed. Knowledge tests were developed to test the soldier on content drawn from the same reference materials. After preliminary versions of the work sample and knowledge measures had been developed, they were administered to approximately five soldiers in each MOS by four non-commissioned officers (NCO). Based on observations from this pilot test, including feedback from the NCO and soldiers, tests were revised and prepared for field testing.

Field Tests. The field tests were designed as a large scale testing of the measures developed to determine what final revisions were needed before the measures were used as criteria for the predictor measures developed in this project. First term soldiers, for whom the tests were developed, participated in the field tests. The minimum length of service for any examinee was 10 months; the maximum length was 36 months. In no occupational specialty did the range of experience, from minimum to maximum length, exceed 17 months. The number of soldiers tested, and the testing locations, are shown in Table 2. Job knowledge tests, consisting of 3 to 16 items per task, were administered by project staff. Work sample administration was supervised by project staff; actual scoring of work sample measures was executed by NCO familiar with the task content and trained in scoring procedures. Number of work sample task steps varied from 4 to 124; only one task exceeded 52 steps.

## Results

### Literature Review

Table 3 shows the results of using meta-analytic procedures to review the findings on correlations between work sample and job knowledge tests. Moderately high correlations were obtained for both types of work samples, with correlations based on verbal work samples exceeding those based on motor work samples. When and only when correlations were corrected both for sample size and attenuation was there evidence that the variability of correlations within work sample category was less than the variability overall.

Table 3 also shows the impact of grouping the investigations which used motor work samples by occupation. Only three of the four major occu-

Table 2

Soldiers by MOS by Location

| LOCATION | MOS[a] | | | | | | | | | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|
| | 11B | 13B | 19E | 31C | 63B | 64C | 71L | 91A | 95B | |
| Fort Hood | | | | | | | 48 | | 42 | 90 |
| Fort Lewis | 29 | | 30 | 16 | 13 | | | 24 | | 112 |
| Fort Polk | 30 | | 31 | 26 | 26 | | 60 | 30 | 42 | 245 |
| Fort Riley | 30 | | 24 | 26 | 29 | | 21 | 34 | 30 | 194 |
| Fort Stewart | 31 | | 30 | 23 | 27 | | | 21 | | 132 |
| USAREUR | 58 | 150 | 57 | 57 | 61 | 155 | | 58 | | 596 |
| TOTAL | 178 | 150 | 172 | 148 | 156 | 155 | 129 | 167 | 114 | 1369 |

Note. Data above are reproduced from "Criterion Reduction and Combination via a Participative Decision Making Panel" by J.P. Campbell and J.H. Harris, 1985. Paper presented at the meeting of the American Psychological Association. Adapted by permission.

[a]11B - Infantryman

13B - Cannon Crewman

19E - Tank Crewman

31C - Radio teletype operator

63B - Vehicle mechanic

64C - Motor transport operator

71L - Administrative Specialist

91A - Medical Specialist

95B - Military Police

[b]Fort Hood, TX; Fort Lewis, WA; Fort Polk, LA; Fort Riley, KS; Fort Stewart, GA; USAREUR = Several military units within Germany

Table 3

Previous Results Classified by Type of Hands-On Test and (for Motor) Type of Occupation

| Investigations | Total Sample Size | $\bar{r}_{kw1}$ | | $\bar{r}_{kw2}$ | |
|---|---|---|---|---|---|
| | | $\bar{r}$ | $SD_p$ | $\bar{r}$ | $SD_p$ |
| All | 4151 | .46 | .12 | .61 | .16 |
| Verbal (Firefighter Supply Specialist, Cartographer, Customs Inspector, Tax Investigator, Claims Examiner) | 1744 | .51 | .10 | .70 | .12 |
| Motor | 2407 | .42 | .13 | .54 | .15 |
| Motor: Operations (Cook, Armor Repairman, Automotive Mechanic, Vehicle Mechanic) | 887 | .45 | .07 | .60 | .11 |
| Motor: Combat/Electronics | 1220 | .36 | .10 | .47 | .15 |
| Combat Arms (Armor Crewman, Infantry Rifleman) | 561 | .47 | 0 | .62 | 0 |
| Electronics (Electronics Maintenaance, Ground Radio Repair, Radar Mechanic, Electronics Technician, Field Radio Repair) | 659 | .27 | 0 | .34 | .05 |
| Motor: Skilled Technical (Medical Laboratory Worker, Radiologic Technologist) | 300 | .58 | .14 | .67 | .09 |

Note. Explanation of terms:

$\bar{r}_{kw1}$ - Correlation between knowledge and work sample test scores corrected for sampling error.

$\bar{r}_{kw2}$ - Correlation between knowledge and work sample test scores corrected for sampling error and attenuation.

$SD_p$ - Adjusted standard deviation of reported investigations.

pational groupings were represented; there were no previous motor work sample investigations which examined clerical occupations. If the combat category is viewed as two separate categories, combat arms and electronics, then the correlational findings are found to be consistently high in every occupation except electronics, and are found to be less variable within specific occupational groupings than within the motor category taken as a whole.

Project A

Table 4 shows means and standard deviations for both work sample and job knowledge tests for each of the Project A occupations. Table 5 shows the correlations obtained in each of the Project A occupations before and after correction for attenuation as well as the reliability coefficients used to make the corrections. Reliability coefficients shown are split half (odd-even).

Table 6 shows overall Project A results and Project A results grouped by occupational category. The variability of correlations within each of the occupational groupings was consistently less than the variability of the correlations across all occupations. The correlations were clearly lowest in the skilled technical category; otherwise, there were no major differences between groupings.

Combining Project A with Earlier Results

Table 7 compares Project A results with those reported in the literature and shows the impact of pooling both sets of results together. Since all Project A work sample tests fit the motor classification, the comparison is based entirely on investigations of the same type.

Table 4

Work Sample and Knowledge Test Means and Standard Deviations: Project A

| Occupation | Sample Size | Knowledge Test | | Work Sample | |
|---|---|---|---|---|---|
| | | M | SD | M | SD |
| Motor Transport Operator | 140 | 60.61 | 10.12 | 72.91 | 9.08 |
| Infantryman | 162 | 54.08 | 11.40 | 56.14 | 12.26 |
| Administrative Specialist | 126 | 59.13 | 10.75 | 62.11 | 9.91 |
| Cannon Crewman | 146 | 59.18 | 12.73 | 54.53 | 13.96 |
| Tank Crewman | 106 | 60.56 | 10.85 | 57.13 | 8.35 |
| Radio Teletype Operator | 127 | 57.76 | 10.59 | 80.06 | 10.74 |
| Vehicle Mechanic | 126 | 64.12 | 9.22 | 79.76 | 8.37 |
| Medical Specialist | 138 | 71.74 | 8.60 | 83.38 | 11.45 |
| Military Police | 112 | 64.70 | 9.01 | 70.77 | 5.75 |

[a]Sample size shown here represents number of soldiers for whom both knowledge test and work sample data were available. Sample size shown in Table 2 represents number of soldiers for whom any data were available.

Table 5

Project A Work Sample and Job Knowledge Correlations

| Occupation | Reliability | | Correlation | |
|---|---|---|---|---|
| | $\underline{r}_{kk}$ | $\underline{r}_{ww}$ | $\underline{r}_{kw1}$ | $\underline{r}_{kw2}$ |
| Motor Transport Operator | 61 | 59 | 59 | 98 |
| Infantryman | 84 | 49 | 55 | 86 |
| Administrative Specialist | 71 | 66 | 52 | 76 |
| Cannon Crewman | 83 | 82 | 41 | 50 |
| Tank Crewman | 76 | 56 | 39 | 60 |
| Radio Teletype Operator | 75 | 44 | 37 | 57 |
| Vehicle Mechanic | 67 | 49 | 31 | 54 |
| Medical Specialist | 80 | 35 | 21 | 40 |
| Military Police | 53 | 30 | 11 | 28 |

Note. See Table 1 for explanation of terms.

Table 6

Project A Results Classified by Type of Occupation

| Category | Total Sample Size | $r_{kw1}$ | | $r_{kw2}$ | |
|---|---|---|---|---|---|
| | | $\bar{r}$ | $SD_r$ | $\bar{r}$ | $SD_r$ |
| Overall | 1183 | .39 | .13 | .62 | .19 |
| Clerical (Administrative Specialist) | 126 | .52 | — | .76 | — |
| Operations (Vehicle Mechanic, Motor Transport Operator, Radio Teletype Operator) | 393 | .43 | .10 | .71 | .16 |
| Combat (Infantryman, Cannon Crewman, Tank Crewman) | 414 | .46 | 0 | .67 | .14 |
| Skilled Technical (Military Police, Medical Specialist) | 250 | .17 | 0 | .35 | 0 |

Note. See Table 3 for explanation of terms.

Table 7

Combining Project A with Other Results (Motor)

| Category | Sample Size | | | $r_{kw1}$ | | | | $r_{kw2}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Previous | Project A | Total | $\bar{r}_o$ | $\bar{r}_a$ | $\bar{r}_c$ | $SD_p$ | $\bar{r}_o$ | $\bar{r}_a$ | $\bar{r}_c$ | $SD_p$ |
| Overall | 2407 | 1183 | 3590 | .42 | .39 | .41 | .13 | .54 | .62 | .57 | .17 |
| Clerical | ---- | 126 | 126 | — | .52 | .52 | — | — | .76 | .76 | — |
| Operations | 887 | 393 | 1280 | .45 | .43 | .45 | .08 | .60 | .71 | .63 | .15 |
| Combat/ | | | | | | | | | | | |
| Electronics | 1220 | 414 | 1634 | .36 | .46 | .39 | .09 | .47 | .67 | .52 | .17 |
| Combat Arms | 561 | 414 | 975 | .47 | .46 | .47 | 0 | .62 | .67 | .64 | .09 |
| Electronics | 659 | --- | 659 | .27 | -- | .27 | 0 | .34 | -- | .34 | .05 |
| Skilled | | | | | | | | | | | |
| Technical | 300 | 250 | 550 | .58 | .17 | .39 | .22 | .67 | .35 | .48 | .17 |

Note. Explanation of new terms (see Table 3 for additional explanation):

$\bar{r}_o$ - Mean correlation from earlier (other) investigations.

$\bar{r}_a$ - Mean correlation from Project A.

$\bar{r}_c$ - Mean correlation from combined set of investigations (Project A and other)

Before correction for attenuation, Project A correlations were somewhat lower than those reported earlier; after correction, Project A correlations were somewhat higher. When individual occupational groupings were considered, the greatest difference between the Project A results and those from the literature was found in the skilled technical category. Here, the correlations found in the literature were relatively high; those found in Project A were exceptionally low. If combat arms and electronics are combined as a single combat category, correlations found in the literature in this category appear relatively low; if they are viewed as separate categories, then the Project A and literature findings with respect to combat arms are highly consistent. Variability by category is not consistently reduced from overall variability, although in no case does variability within a category exceed overall variability.

## Discussion

### Summary

The results, both of earlier investigations and of Project A, showed a reasonably high relationship between work sample and job knowledge measures. The relationship was consistently positive, but the magnitude of the positive relationship was sufficiently variable to merit examination of possible moderator variables. The correlations for investigations using verbal work samples tended to exceed those using motor work samples, suggesting that one should use caution in combining results from the two types of research. Comparisons here have thus been made only within the same work sample category.

The occupational categories explored in this effort are not presumed to accurately represent the job factors which may influence the relation-

ship between work sample and job knowledge tests. However, it was hoped that they might represent a first step toward identifying such factors. In this sense, the effort seems to have met with some success. Use of the categories did tend to reduce the dispersion of reported correlational values, particularly when combat arms and electronics were treated as separate categories. The distinction between operations and combat arms categories seemed to have had little effect upon these correlations. The results for the skilled technical category in Project A diverged sharply from previously reported results in this category, a point that will receive more discussion shortly. Examination of reported findings in the electronics category indicated that correlations there have been consistently lower than in any category considered in this effort.

These trends must be interpreted with some caution. The number of total investigations considered was not large; the number within any particular occupational grouping was considerably smaller. The adjustments applied to the correlational coefficients are not purported to have totally eliminated error; the adjusted correlations are still at best a rough approximation of true correlational values. Uncontrolled differences in tests, samples and occupations within each grouping may have biased these results in some undetermined way.

Given these caveats, it is still of interest to explore what characteristics of electronics maintenance jobs might contribute to lower correlations between work sample and knowledge tests in this occupational grouping than in other groupings. A review of the descriptions of the work sample tests in the electronics jobs revealed that one thing they all had in common was a troubleshooting component which, in most cases, was

the major or only part of the test. A typical troubleshooting task would require the examinee to identify a malfunction; the examinee might then be required to take corrective action. Essentially, the examinee was required to exhibit problem solving behavior in an equipment-intensive environment. Apparently, the ability to perform well on such a task was not highly related to the knowledge tested on a comparable written test. Perhaps no written test can tap such an ability particularly well.

The correlational findings in the electronics category are particularly interesting when considered in the context of Vineberg and Taylor's (1972b) thesis that correlations would be lower in highly skilled than in relatively unskilled occupations. Vineberg and Taylor (1972b) have already analyzed selected occupations in the combat arms (armor crewman), operations (cook, armor repairman), and clerical (supply specialist) occupations and determined that the skill requirements in these occupations were low. A systematic analysis of the skill requirements in representative occupations in the electronics category would appear to be an appropriate next step. The proposition that this occupation may have greater skill requirements than clerical, operations or combat occupations and that such skill requirements may limit correlations between work sample and job knowledge measures is both sufficiently intriguing and sufficiently plausible to merit direct investigation.

Project A Contributions

What are the contributions from Project A to the overall literature on relationships between work sample and job knowledge tests? Thus far, Project A data has reinforced earlier findings showing high positive relationships between work sample and job knowledge tests overall and in most

occupational categories. It has strengthened the evidence showing variability across all findings and showing that such variability may be somewhat reduced when findings are classified by occupational category. It has identified the skilled technical category as one for special study because of some unexpectedly low relationships observed in this occupational grouping.

Those are the contributions thus far. When Project A is complete, it will have provided substantially greater stability to any judgments that are made regarding relationships between job knowledge and work sample tests. Given current plans and barring unforeseen circumstances, the majority of the examinees on whom data on relationships between the two types of measures are at that time available will have been tested in Project A.

Implications for Project A

What are the implications of these findings for Project A? The results from the literature define a context in which Project A findings might be evaluated. Consistency with past findings, while not necessarily a guarantee that Project A measures were competently developed and tested, does at least diminish the likelihood that Project A developmental procedures went seriously astray from accepted practice. Divergence from past findings would not necessarily demonstrate deficiencies in Project A measures but would suggest the need for explanation.

In general, the pattern of correlations found in Project A was similar to the pattern found in previous investigations. The discrepant findings in the skilled technical occupational category clearly need some explanation. Also, the differential impact of the adjustment for

attenuation requires examination. These issues we shall address now.

Skilled Technical Findings. Why should the Project A results in the skilled technical occupational grouping have been particularly discrepant from previous findings obtained in this category? Do the results suggest aberrations in Project A test developmental procedures with respect to the relevant occupations? Since the same developmental procedures were followed for these occupations as for others in which no major discrepancies were apparent, the explanation would appear to be elsewhere.

It was noted earlier that restriction in range may have differentially affected correlational findings, although sufficient data to correct for such restriction was not available in the investigations reviewed. Examination of the samples tested in Project A relative to those tested in earlier investigations involving skilled technical occupations does suggest that differential selection standards had been applied. All examinees in Project A had to achieve qualifying scores on the Armed Forces Qualifying Test (AFQT) and one or more aptitude composites. Both the AFQT and the aptitude composites are composed of subtests from the Armed Services Vocational Aptitude Battery, a general cognitive test. Screening processes produced means on the AFQT that were higher in the Project A skilled technical specialties, military police ($M$ = 64.80) and medical specialist ($M$ = 59.61), than in any of the other Project A occupations examined.

Comparable data was not available on examinees in the two investigations from the literature review that involved occupations classified as skilled technical, but the information that was available did not suggest that the subjects sampled were either highly selected or

particularly homogeneous in terms of ability. In the Schoon (1979) project, medical technologists tested had widely varying backgrounds and current working assignments, sharing only the distinction of meeting eligibility requirements for admission to a specified proficiency examination. The authors reported substantially higher standard deviations on the work sample for these individuals than for a pre-test group of individuals who had graduated from an accredited medical technology program. In the Livingston (1978; Education Testing Service, 1977) effort, both credentialed and non-credentialed radiologic technologists were included in the total sample. When the results pertaining only to credentialed radiologists were examined, the uncorrected correlation between work sample and knowledge tests dropped from .42 for the overall sample to 0 for the restricted sample.

Attenuation Adjustment. The disproportionately large impact of the adjustment for attenuation upon Project A correlations simply reflects the fact that the Project A reliabilities used in the adjustment for attenuation were lower than the reliabilities generally found in the literature. Work sample reliabilities found in the literature, weighted according to sample size, averaged .71, compared to .53 for the work samples used in Project A. Job knowledge test reliabilities reported in the literature averaged .83; job knowledge reliabilities used for Project A attenuation corrections averaged .73. It has already been suggested that internal consistency reliability estimates may have some limitations, particularly with respect to work samples. Nevertheless, the question remains, why were the Project A reliability estimates shown here below those typically found?

Part of the answer is that, while estimates reported in the literature were based on complete tests, Project A results were not. The comparison between work sample and job knowledge measures in Project A was based on only 15 tasks. Job knowledge tests were also developed for an additional 15 tasks. Ultimately, a 30-task test will be generated based on the 15 tasks for which both types of measures have been developed and the 15 tasks for which only job knowledge tests have been developed. A reliability estimate based on the 15 tasks used for comparing work sample and job knowledge measures was the appropriate basis for correcting correlations generated from this comparison. However, a better estimate of the reliability of the 30 task job knowledge tests, obtained using the Spearman Brown formula, is .84 for the job knowledge tests, relative to the .83 value reported in the literature. Using the same formula, it can be estimated that the reliability of a 30-task work sample test would have been .69, relative to the .71 value reported in the literature.

A second factor to consider in comparing reliability estimates observed in Project A with those reported in other investigations is the nature of the tests on which those estimates were based. Two different strategies were available in developing these tests. Developers could, on the one hand, attempt to measure every major job dimension. This was the strategy adopted in Project A. Clusters of tasks were identified and tests developed to measure one or more tasks in each such cluster. The approach tends to maximize job coverage but also to produce relatively heterogeneous tests and depressed estimates of internal consistency. The other approach would be to build tests around a relatively few job dimensions that are judged to be central to a given job. This is the approach

that was typically taken in the investigations reviewed here. For example, Trattner, et al. (1977) built an entire work sample for claims examiners around a standardized claim to be adjudicated by the examinee. Based on the examinee's performance, scores were generated in five duty areas. This strategy tends to yield a more homogeneous test and, accordingly, a higher estimate of internal consistency than the Project A approach.

In the skilled technical category, a third factor which may have differentially affected Project A reliability estimates and those from other research efforts is restriction in range. The basis for suspecting differential restriction in range in this occupational category has already been discussed. It is noteworthy that the lowest Project A work sample and job knowledge test reliabilities were found in a skilled technical occupation, military police.

The final factor to be considered in comparing Project A reliability estimates with those reported in the literature is that the Project A estimates were based on draft measures while estimates reported in the literature were based on the final versions of the measures developed. Revisions have since been made to the Project A measures based on the field test data; the reliability of the revised measures remains to be determined.

Summary. We have examined the relatively few discrepancies between Project A results and findings from earlier investigations in an attempt to determine why these discrepancies might have occurred. This examination did not suggest any serious deficiencies in Project A measures, but we have not been complacent about these measures. Further refinements, to be briefly described shortly, have been made.

What other implications do these findings have for Project A? Critical to the final development of Project A criterion composites is the question of interchangeability of work sample and job knowledge tests. If these two types of measures are perfectly related, then one or the other is superfluous. In general, the Project A results in conjunction with the results of other investigations suggest that that is not a realistic expectation. Within most occupational groupings, reasonably high correlations are possible, but enough variance remains unaccounted for to indicate that the measures are not totally interchangeable. Interestingly, Project A did generate one uncorrected correlation, for motor transport operator, that was sufficiently close to the reliabilities of the knowledge and work sample measures to suggest the possibility of interchangeability. However, it would be placing too much confidence on both the reliability estimates and the correction using such estimates to firmly assert at this time there was enough evidence to use either test alone and drop the other.

The question of how to optimally combine Project A work sample and job knowledge tests has not yet been finally resolved. The next stage in Project A, involving concurrent administration of predictor and criterion measures, has already begun. Based on presently available data, it seems imprudent to reduce either work sample or job knowledge tests any more than necessary. Thus, the strategy in terms of additional development work for the concurrent validation has been to retain the best aspects of both types of measures to the extent possible. Work sample tests have been generally maintained in their entirety. Job knowledge tests, which occupied a four-hour time block for the field tests but must fit into a

two-hour block for the concurrent validation, have been reduced by deleting those items which appear to be most expendable, whether because of low discriminability, low reliability, or some other negative indicator or combination of indicators. Only very infrequently have tests of entire tasks been deleted. Thus, the concurrent validation will offer an additional opportunity to examine the relationship between job knowledge and work sample measures, with the advantages of a larger subject pool and more refined job knowledge tests but with the disadvantage that the knowledge tests will be shorter. When the concurrent validation data has been collected, more precise conclusions about the appropriate mix of work sample and job knowledge tests in a criterion composite will be drawn.

References

Asher, J. J. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.

Brown, G. H. Zaynor, W. C., Bernstein, A. J., & Schoemaker, H. A. (1959). Development and evaluation of an improved field radio repair course (Tech. Rep. No. 58). Washington, DC: Human Resources Research Office, George Washington University.

Campbell, J. C., School and job performance measurement (1984). In N. K. Eaton, M. H. Goer, J. H. Harris & L. M. Zook (eds.), Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year (Tech. Rep. 660). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). An investigation of sources of bias in the prediction of job performance: A six year study. Princeton, NJ: Educational Testing Services.

Corts, J. B., Muldrow, T. W., & Outerbridge, A. M. (1977). Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job performance for customs inspectors. Washington, DC: Personnel Research and Development Center, U.S. Office of Personnel Management.

Crowder, N. A., Morrison, E. J., & Demaree, R. G. (1954). Proficiency of Q-24 radar mechanics: VI Analysis of intercorrelations of measures (Tech. Rep. No. 54-127). San Antonio, TX: Air Force Personnel and Training Research Center.

Educational Testing Service (1977). Validation of the proficiency examination for diagnostic radiologic technology. Princeton, NJ: Educational Testing Service.

Engel, J. G., & Rehder, R. J. (1970). A comparison of correlated-job and work-sample measures for general vehicle repairman (Tech. Rep. No. 70-16). Fort Knox, KY: Human Resources Research Organization.

Evans, R. M., & Smith, L. J. (1953). A study of trouble shooting ability on electronic equipment. Urbana, IL: College of Education, University of Illinois.

Foley, J. P. (1974). Evaluating maintenance performance: An analysis (Tech. Rep. No. 74-57). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Foley, J. P. (1977). Performance measurement of maintenance (Tech. Rep. No. 77-76). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Grings, W. W. (1953). A methodological study of electronics troubleshooting skills: II. Inter-comparisons of the MASTS test, a job sample test, and ten reference tests administered to fleet electronics technicians (Electronics Personnel Research Group Rep. No. 10). Los Angeles: Department of Psychology, University of Illinois.

Human Resources Research Organization and American Institutes for Research, The Task 5 Staff of Project A (1984). Selecting job tasks for criterion tests of MOS proficiency (Working Paper 84-25). Alexandria, VA: Selection and Classification Technical Area, U.S.

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

Hunter, J. E., Schmidt, F. S., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills: 1982.

Livingston, S. A. (1978). Concurrent validation of a written proficiency examination. Paper presented at the meeting of the American Educational Research Association, Toronto.

Maier, M. H., & Hiatt, C. M. (1982). Evaluating measures of job performance in three Marine Corps skills (Research Memorandum 82-3153). Alexandria, VA: Center for Naval Analyses.

O'Leary, B. S., & Trattner, M. H. (1977). Research base for the written portion of the Professional and administrative career examination (PACE): Prediction of job performance for internal revenue officers. Washington, DC: Personnel Research and Development Center, U.S. Office of Personnel Management.

Osborn, W. C. (1984). A summary of the development procedures for job performance measurement in project A (Working Paper 84-20). Alexandria, VA: Selection and Classification Technical Area, U.S. Army Research Institute for the Behavioral and Social Sciences.

Osborn, W., & Ford, P. (1977). Knowledge tests of manual task procedures. Paper presented at the meeting of the Military Testing Association, San Antonio, TX.

Rosse, R. L., Borman, W. C., Campbell, C. H., & Osborn, W. C. (1983). Grouping Army occupational specialties by judged similarity. Paper presented at the meeting of the Military Testing Association, Gulf Shores, AL.

Saupe, J. L. (1955). An analysis of trouble-shooting behavior of radio mechanic tranees (Research Report 55-47). Lackland Air Force Base, TX: Air Force Personnel and Training Center.

Schoon, C. G. (1979). Correlation of performance on clinical laboratory proficiency examinations with performance in clinical laboratory practice. New York: Professional Examination Service, 1979.

Trattner, M. H., Corts, D. B., Van Rijn, P. P., & Outerbridge, A. M. (1977). Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job performance for claims authorizers in the social insurance claims examining occupation. Washington, DC: Personnel Research and Development Center, U.S. Office of Personnel Management.

van Rijn, P. & Payne, S. S. (1980). Criterion related validity research base for the DC firefighter selection test. Washington, DC: Personnel Research and Development Center, U.S. Office of Personnel Management.

Vineberg, R., & Taylor, E. N. (1972a). Performance in four army jobs by men at different aptitude levels: 3. The relationship of AFQT and job experience to job performance (Tech. Rep. No. 72023). Alexandria, VA: Human Resources Research Organization.

Vineberg, R., & Taylor, E. N. (1972b). Performance in four army jobs by men at different aptitude (AFQT) levels: 4. Relationships between performance criteria (Tech. Rep. 72-23). 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.

Vineberg, R., & Taylor, E. N., & Sticht, T. G. (1970). Performance in five army jobs at different aptitude (AFQT) levels: 2. Development and description of instruments (Tech. Rep. No. 70-20). Monterey, CA: Human Resources Research Organization.

Williams, W. L., Jr., & Whitmore, P. D., Jr. (1959). The development and use of a performance test as a basis for comparing technicians with and without field experience: The NIKE AJAX IFC maintenance technician (Tech. Rep. 52). Washington, DC: Human Resources Research Office, George Washington University.

Wing, H., Peterson, N. G., & Hoffman, R. G. (1984). Expert judgements of predictor-criterion validity relationships. Paper presented at the meeting of the American Psychological Association, Toronto.

# DEVELOPMENT OF COGNITIVE/PERCEPTUAL MEASURES: SUPPLEMENTING THE ASVAB

Jody L. Toquam, Marvin D. Dunnette, VyVy Corpe,
Jeffrey J. McHenry, Margaret A. Keyes, Matthew K. McGue,
Janis S. Houston, Teresa L. Russell, and Mary Ann Hansen
Personnel Decisions Research Institute

August 1985

Paper presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

## Author Notes

This paper was prepared as part of a symposium on "Expanding the Measurement of Predictor Space for Military Enlisted Jobs," presented at the annual meeting of the American Psychological Association, August, 1985. Each of the papers discusses a different aspect of developing a set of predictor measures for the Army's Project A, an effort designed to improve the selection, classification and utilization of enlisted personnel. 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 is being funded by the U.S. 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 or policies of the Army Research Institute or the Department of the Army.

# DEVELOPMENT OF COGNITIVE/PERCEPTUAL MEASURES: SUPPLEMENTING THE ASVAB

## Introduction

The purpose of this paper is to (1) summarize the activities surrounding the development of cognitive ability measures that supplement or provide information about Army applicants' abilities not currently tapped by the Armed Services Vocational Aptitude Battery, or ASVAB; (2) describe the measures developed and the constructs they are designed to tap; (3) describe the psychometric qualities and characteristics of the new measures; and (4) report decisions made about which measures to include in the Trial Battery (administered during Summer and Fall 1985).

Before describing the new tests, we first examine the content of the current military selection and classification battery, the ASVAB, and then provide a brief review of the process involved in identifying the constructs for inclusion in the Pilot Trial Battery. As noted above, the purpose for designing these new measures is to assess those cognitive abilities not currently tapped by the ASVAB. Briefly, this battery contains ten subtests. Scores on four of these are used to calculate the Armed Forces Qualification Test (AFQT) score for selection purposes. Scores on the ten subtests are used in different combinations to determine applicants' qualifications for different military occupational specialties (MOS). Drawing from results of a factor analysis of the ASVAB , the battery

assesses verbal ability, speeded performance, quantitative ability, and technical knowledge (Kass, Mitchell, Grafton, & Wing, 1982).

As previously described (Peterson, 1985), the process of identifying and developing new measures to supplement the ASVAB began with a review of the cognitive abilities domain. This included a literature search through available computerized data bases, current journals, military bibliographies, test manuals and so on to locate information about cognitive ability constructs. This information was used to impose structure on the cognitive ability domain (i.e. establish a cognitive/perceptual abilities taxonomy) and then to examine and summarize validity data for the different types of ability constructs.

Next, the resulting cognitive ability taxonomy was used to identify important constructs for inclusion in the expert judgment task. Again, the constructs included in that activity and resulting validity estimates have been described by Peterson (1985). These data were then used to prioritize our measurement development efforts for the Pilot Trial Battery. (The Pilot Trial Battery is the term used for the battery of experimental tests administered at Fort Carson, Fort Campbell, Fort Lewis, and Fort Knox. This battery includes twelve paper-and-pencil measures, ten cognitive and two non-cognitive, and ten computerized measures, seven cognitive perceputal and three psychomotor.) A copy of the prioritized list of constructs is provided on Table 1.

Table 1

Cognitive/Perceptual Construct Prioritized Test Development List[a]

| | Priority | Cognitive Construct |
|---|---|---|
| Paper-and-Pencil | 1 | Spatial Visualization - Rotation & Scanning |
| | 2 | Spatial Visualization - Field Independence |
| | 3 | Spatial Orientation |
| | 4 | Induction - Figural Reasoning |
| Computer | 5 | Reaction Time - Processing Efficiency |
| | 6 | Memory - Number Operations |
| | 7 | Memory - Short Term Memory |
| | 8 | Perceptual Speed and Accuracy |

[a]Note that Movement Judgment was not included in Prioritized Construct List.

## Determining the Method of Administration

After identifying the target constructs for test development purposes, we then focused on determining the method of administration for each. Rosse (1985) has noted the various constraints and issues surrounding the development of computerized tests. Factors impacting on the decision to develop paper-and-pencil versus computerized tests include the following:

- Administration requirements of target constructs. For example, to assess reaction time on simple or more complex tasks adequately, computer administration is a necessity. Hence, those constructs that involve a reaction time component, such as Simple and Choice Reaction Time, were slated for computerized administration.
- Ease of adapting paper-and-pencil measures to the computer. For example, some test items such as the space visualization or figural reasoning items pose problems in transferring the graphics from paper-and-pencil format to the computer.
- All computerized tests must be self-administering. Thus, for each computerized test, instructions must clearly explain the required task while ensuring that verbal ability or reading level requirements remain at a moderately low level.
- Practical test administration issues. Test administration time as well as equipment availability also guided the selection of measures for computer administration. Because

of these limitations, not all cognitive tests could be administered via computer. Hence, we identified measures of constructs that by definition require computer administration and measures of constructs that address basic research questions about computer-administered tests.

Given the above considerations, we determined that measures of spatial visualization, spatial orientation, and induction would be assessed via paper-and-pencil. Measures of the remaining four constructs (as seen in Table 1) would be assessed via computer. These include measures of Reaction Time, Perceptual Speed and Accuracy, Number Memory, Short Term Memory, and Movement Judgment.

Because issues related to test development vary somewhat for the two modes of test administration, we first describe the new paper-and-pencil measures and then describe the computer tests designed for the Trial Battery. We begin with a description of the target constructs for the paper-and-pencil measures and focus on the issues in test development.

Paper-and-Pencil Measures: Construct and Test Descriptions

Table 2 contains a list and description of the newly developed paper-and-pencil measures included in the Pilot Trial Battery. Note that the tests are listed by construct. The constructs and target criterion performance behaviors predicted by each are described briefly below.

Spatial Visualization--Rotation

This involves the ability to mentally restructure or manipulate parts of a two- or three-dimensional figure. It serves as a potentially effective predictor of success in MOS that involve mechanical operations, construction and drawing or using maps. The two tests developed to measure this construct include Assembling Objects and Object Rotation.

Spatial Visualization--Scanning

This includes the ability to visually survey a complex field and to find a pathway through it. According to our expert judges, measures of this construct are potentially effective as predictors of success for Army MOS involving electrical or electronics operations, using maps in the field, and controlling air traffic. The two measures designed to assess this construct include the Path Test and the Maze Test.

Spatial Visualization--Field Independence

This includes the ability to find a simple form when it is hidden in a complex pattern. This type of measure is expected to predict success in MOS that involve detecting and identifying targets, using maps in the field, planning placement of tactical positions, air traffic control and troubleshooting operating systems. The Shapes Test was developed to measure this construct.

Spatial Orientation

This represents the ability to maintain one's bearing with respect to points on a compass and to maintain appreciation of

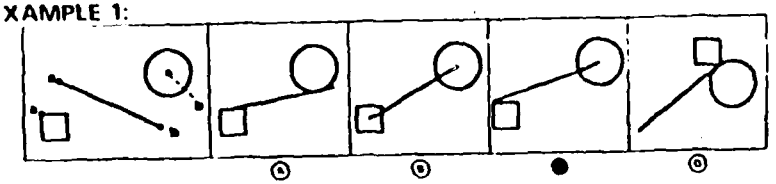
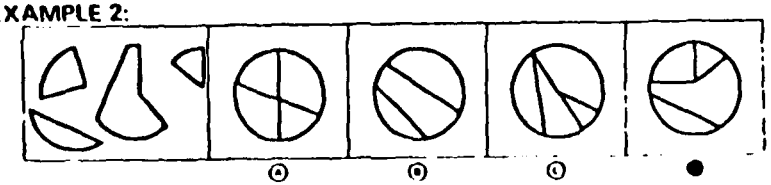
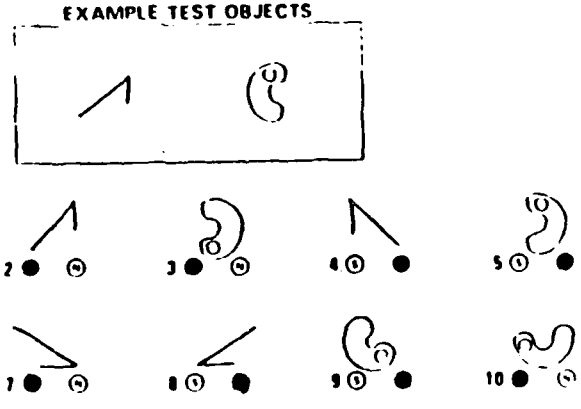
one's location relative to landmarks in the environment. Conceptualization of this construct first appeared during World War II in the Army Air Forces Aviation Psychology Program (Guilford & Lacey, 1947). Dr. Lloyd Humphreys, of the Scientific Advisory Group for Project A, is particularly responsible for emphasizing the usefulness of this construct to us. Based on job observations that we conducted in the field, measures of this construct are expected to be effective predictors of success in a wide variety of MOS, especially those combat MOS that include critical job requirements of maintaining directional orientation using features or landmarks in the environment. Three tests involving different orientation tasks were developed to assess this construct.

Induction--Figural Reasoning

This involves the ability to generate hypotheses about principles governing relationships among several objects. According to the panel of experts, measures of this construct are effective predictors of success in MOS involving troubleshooting, inspecting, and repairing electrical, mechanical or electronic systems, analyzing data, controlling air traffic, and detecting and identifying targets. Two tests involving different task requirements were constructed to assess this construct.

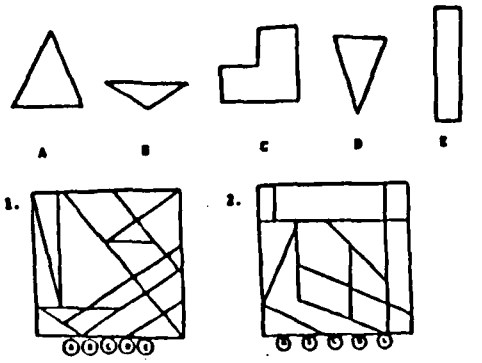
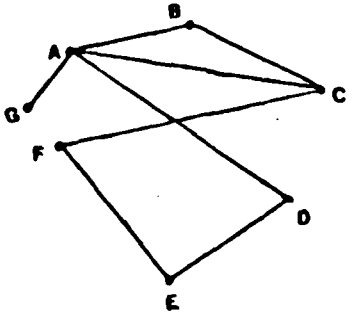
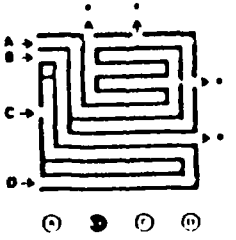
Table 2

Description of Cognitive Paper-and-Pencil Measures

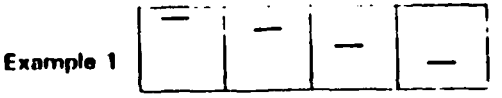
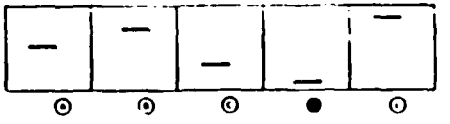
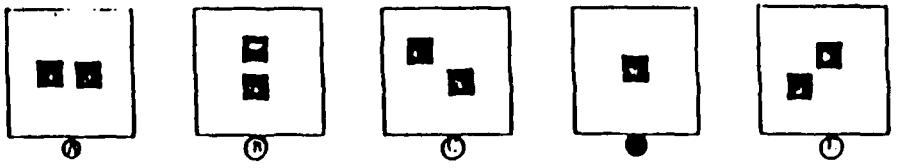
| CONSTRUCT/MEASURE | DESCRIPTION OF TEST | SAMPLE ITEM |
|---|---|---|
| **SPATIAL VISUALIZATION - ROTATION** | | |
| Assembling Objects | The test contains 40 items with a 16 minute time limit. The subject's task involves figuring out how an object will look when its parts are put back together again. There are two types of problems in the test. In one part, the item shows a picture of labelled parts. By matching the letters, it can be "seen" where the parts should touch when the object is put together correctly. The second type of problem does not label any of the parts. The parts fit together like the pieces of a puzzle. In each section, four possible figures are provided and the subject must pick the correct one. | EXAMPLE 1:  EXAMPLE 2:  |
| Object Rotation | The test contains 90 items with a 7 1/2 minute time limit. The subject's task involves examining a test object and determining whether the figure represented in each item is the same as the object, only rotated, or is not the same as the test object (e.g., flipped over). For each test object there are 5 test items, each requiring a response of "same" or "not same". | EXAMPLE TEST OBJECTS  |

/continued

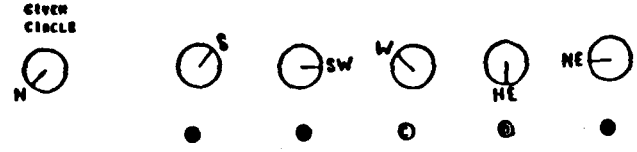
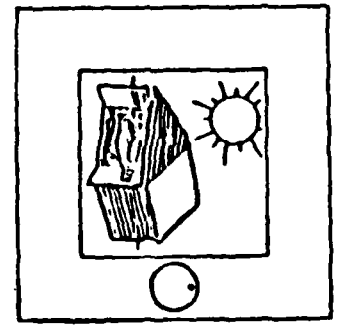
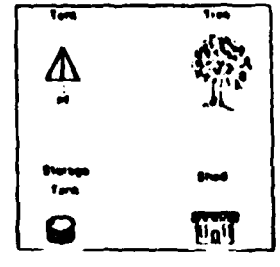
Cognitive Pencil-and-Paper Measures

| CONSTRUCT/MEASURE | DESCRIPTION OF TEST | SAMPLE ITEM |
|---|---|---|
| **SPATIAL VISUALIZATION - FIELD INDEPENDENCE** | | |
| Shapes | The test contains 54 items with a 16-minute time limit. At the top of each test page are five simple shapes; below these shapes are six complex figures. Subjects are instructed to examine the simple shapes and then to find the one simple shape located in each complex figure. |   |
| **SPATIAL VISUALIZATION - SCANNING** | | |
| Path | The test contains 44 items with an 8-minute time limit. Subjects are required to determine the best path or route between two points. Subjects are presented with a map of airline routes or flight paths. The subject's task is to find the "best" path or the path between two points that requires the fewest number of stops. |   The route from: Number of Stops: 1. A to F ❶ ② ③ ④ ⑤ 2. G to E ① ② ③ ④ ⑤ 3. C to D ① ② ③ ④ ⑤ 4. G to F ① ② ③ ④ ⑤ |
| Mazes | The test contains 24 items with a 5 1/2 minute time limit. Each item is a rectangular maze with four labelled entrance points and four exit points. The task is to determine which of the four entrances leads to a pathway through the maze and to one of the exit points. |   ⑥ ⑨ ⑦ ⑩ |

/continued

| CONSTRUCT/MEASURE | DESCRIPTION OF TEST | SAMPLE ITEM |
|---|---|---|
| INDUCTION | | |
| Reasoning 1 | The test contains 30 items with a 12 minute time limit. Subjects are presented with a series of four figures. The task is to identify the pattern or relationship among the figures and then to identify from among five possible answers the one figure that appears next in the series. | FIGURE SERIES<br>Example 1<br>  <br>POSSIBLE ANSWERS<br>  |
| Reasoning 2 | The test contains 32 items with a 10 minute time limit. Subjects are presented with five figures. They are then asked to determine which of the four figures are similar in some way, thereby identifying the one figure that differs from the others. | Example 1:<br>  |

/continued

| CONSTRUCT/MEASURE | DESCRIPTION OF TEST | SAMPLE ITEM |
|---|---|---|
| **SPATIAL ORIENTATION** | | |
| Orientation 1 | The test contains 150 items (30 5-item sets) with a 10 minute time limit. Each set presents subjects with 6 circles. The first, the Given Circle, indicates the compass direction for North. For most items, North is rotated out of its conventional position (e.g., the top of the circle does not necessarily represent North). Compass directions also appear on the remaining five circles. The subject's task is to determine for each circle, whether or not the direction indicated is correctly positioned by comparing it to the direction of North in the Given Circle. |   |
| Orientation 2 | The test contains 24 items with an 10 minute time limit. Each item contains a picture within a circular or rectangular frame. The bottom of the frame has a circle with a dot inside it. The picture or scene is not in an upright position. The task is to mentally rotate the frame so that the bottom of the frame is positioned at the bottom of the picture. After doing so, the subject must then decide where the dot will appear in the circle. |  |
| Orientation 3 | The test contains 20 items with a 12 minute time limit. Subjects are presented with a map that includes various landmarks such as a barracks, a campsite, a forest, a lake, and so on. Within each item, subjects are provided with compass directions by indicating the direction of one landmark to another, such as "the forest is North of the camp site". Subjects are also informed of their present location relative to another landmark. Given this information, the subject must determine which direction to go to reach yet another structure or landmark. |  1. The shed is due north of the tree. You are at the storage tank. Which direction must you travel to reach the tent? |

## Issues in Test Development

### Target Population

The population for these tests is the same one to which the Army applies the ASVAB (i.e. persons making application to enlist in the U.S. Army). This is, very generally speaking, a population made up of predominantly recent high school graduates, not entering college, from all geographic sections of the United States.

Another point to make about the target population is that it was, practically speaking, inaccessible to us during our development process. We were constrained to using enlisted soldiers to try out the newly developed tests. Enlisted soldiers, of course, represent a restricted sample of the target population in that they all have passed enlistment standards and, furthermore, we could expect that almost all of the first-term soldiers included in the tryouts would also have passed basic and advanced individual training. Thus, the persons used to assess the psychometric quality of the new tests would be presumably more qualified, more able, more persevering, etc. on the average than are the persons in the target population.

The information about the target population and the sample available for the pilot test leads to two major implications that served as general guidelines for our development and pilot testing activites:

(1) The tests to be developed will be applied to a population with a large range of abilities. Therefore, we attempted to develop tests with a broad range of item difficulties. Highly peaked tests were not our goal (e.g. test items with difficulty levels near a certain value such as .50).

(2) The first-term soldiers upon which the tests would be initially tried out are generally higher in ability than the target population. Therefore, the overall difficulty level of the tests should be somewhat higher (i.e. the test should be somewhat easier) than what it would have been if we had had access to an unrestricted sample of the target population.

Power vs. Speed

Another decision to be made about each test was its placement on the power vs. speed continuum. Most psychometricians would agree that a "pure" power test is a test administered such that all persons taking the test are allowed enough time to attempt all items on the test, and that a "pure" speeded test is a test administered such that no one taking the test has enough time to attempt all of the items. In practice, there appears to be a power/speed continuum, and most tests fall somewhere between the two extremes on this continuum.

The decision to develop a power or speeded test or a combination of the two depends in large part on one's definition of the target construct. That is, for a particular test,

definition of the construct may indicate the importance of a speed component. Therefore, during the preliminary test development stage, we categorized each test as a power test, speeded test, or combination of the two using our construct definitions. For example, based on our definition of the construct induction, we designed the test items to represent a very wide range of difficulty levels and established a generous time limit such that most subjects would have time to complete all items. Thus, measures of induction were designed to fall on the power end of the continuum. Our plan for the spatial visualization measures differed from this in that all items were constructed to be moderately easy but a more restrictive time limit was imposed. Thus, these measures were intended to fall toward the speeded end of the continuum.

As a matter of practical definition for this developmental effort, we used an "85% completion" rule-of-thumb to define a power test. That is, if a test could be completed by 85% of all those taking the test, then we considered it a "power" test. Tests with completion rates lower than this were considered to have some "speededness" determining performance on the test, (for example, we somewhat arbitrarily labeled tests with a 70 to 84% completion rate as moderately speeded, and those with completion rates lower that 69% as speeded). This rule-of-thumb, then, served to refine and modify the new tests after each tryout.

## Procedures for Evaluating the Paper-and-Pencil Tests

To evaluate the new measures, four pilot test or tryout sessions were conducted at Fort Carson, Fort Campbell, Fort Lewis, and Fort Knox. In the first tryout at Fort Carson, about 38 soldiers completed each paper-and-pencil test. The number at Fort Campbell was 57 and at Fort Lewis it was 118. At Fort Knox the numbers were 2?0 for time one, and 97-126 for time two, respectively. Procedures for evaluating each test at one or more of these tryouts include the following: construct validity, test item characteristics, internal consistency, and stability.

### Construct Validity

In the course of developing the new paper-and-pencil tests, we examined several published tests designed to tap the same or similar constructs. These published tests served as models for the new tests during the item construction stage. That is, we examined the required task in each published test and reviewed test manuals to ascertain the target population for whom the test was constructed, test item difficuly levels and so on. It is important to note that even though we relied on published tests to serve as models, the new tests differed from the models in terms of the task required and level of item difficulty.

Although differences between the published tests serving as models and the new tests were expected, we wanted to ensure that the new test captured the essence of the target construct.

Therefore, in the first three tryouts, we included several of the "model" tests to assist in the evaluation of the new tests. Commercially published measures of spatial visualization--rotation and scanning, field independence, and induction were included in one or more tryouts. Note that no measures of spatial orientation were included in these tryouts. This is because there are no commercially published tests that measure spatial orientation in a way similar to the way we have defined it here.

Test Item Characteristics

For all tests administered at all tryouts, we examined item difficulty levels and item-total correlations. These data were used to modify test items and to adjust time limitations.

Internal Consistency

This included administering each test as two separately timed halves and then computing the correlation between part one and part two for each test. The Spearman-Brown correction procedure was then used to estimate the reliability for the test as a whole. This procedure was used in addition to computing the Hoyt reliablity because for some of the more highly speeded tests, the Hoyt provides an overestimate of the reliability. (Note: split-half forms of each test were administered at the first three tryout sites--Fort Carson, Fort Campbell, and Fort Lewis.)

Stability

In the last tryout conducted at Fort Knox, we collected test-retest information on a sample of about 100 first-term soldiers. A period of two weeks separated the two test sessions.

### General Findings: Paper-and-Pencil Measures

Below we present some general findings for the paper-and-pencil tests as a whole. The discussion is organized around the four topic areas described above.

Construct Validity

Very few of the newly developed tests correlated above .65 with the designated marker test; most correlations between new measures and marker tests fell between .45 and .60. These values were as expected, given the differences in task requirements and in item difficulty levels between the new test and the marker tests. For example, because of fairly low item difficulty levels (i.e., items were, on the average, very difficult) and restrictive time limitations on some of the marker tests, subjects completed only a small proportion of the items, resulting in highly restricted test scores, thereby reducing the correlation with the new test.

Basically this information suggested to us that although the tests did not duplicate their respective marker tests, they captured the essence of the target construct.

Test Characteristics

In the very first tryout, we discovered that most tests required modification. That is, on some of the tests, the average subject completed all the items and obtained a very high total test score. Analysis of item difficulties and item-total correlations led to modifications of these tests after the Fort Carson tryout. For example, for Assembling Objects, Object Rotation, Orientation I, and Path Test, new items were constructed and added with only minor changes in time limits to increase overall item difficulty levels, and to reduce the possibility of ceiling effects. For the Shapes Test and Maze Test, the items themselves were modified to increase item difficulty.

The reverse situation appeared on the Orientation 2 Test. That is, item difficulty levels were low (mean = .48; the test appeared more difficult than desired). Thus, we examined the difficulty levels across all items to identify the components that resulted in easier and more difficult items. This information was used to construct four additional items of lower difficulty. Further, the time limit was lengthened to ensure that all or nearly all subjects would complete the test.

For the remaining measures, Orientation 3, Reasoning 1 and Reasoning 2, very few changes were required. For example, for a few items in each test, item analysis data revealed that item

total-correlations were higher for a distractor than for the correct response. These items were modified or replaced.

Following each tryout, item difficulty levels as well as item-total correlations were examined to determine what changes were required to improve the test. Subsequent pilot tests or tryouts indicated that the tests, in general, required only minor modifications.

Internal Consisitency

As reported above, split-half reliability estimates were computed for each test following the first three tryouts. Results from the Fort Lewis tryout are presented in Table 3. Note that for all tests, with the exception of Reasoning 2, these values are at acceptable levels, ranging from the high 70's to low 90's.

Also, in the same table are internal consistency estimates computed for each test using the data collected at Fort Knox. Note that these values have been computed using the Hoyt formula and may represent overestimates for some of the more highly speeded tests. Again, with the exception of Reasoning 2, these values range from the low 80's to high 90's.

Stability

Table 3 also contains the test-retest reliability estimates computed for a sample of about 100 first term recruits. These values are lower than the internal consistency estimates, but are at acceptable levels, ranging from .57 to .84. Note that

Table 3

Reliability Estimates for the Ten Paper-and-Pencil Tests

| Test | No. Items | Time Allotted (in minutes) | FORT LEWIS $r_{xx}$ Split Half N = 118 | FORT KNOX $r_{xx}$ Alpha N = 290 | FORT KNOX $r_{xx}$ Test-retest (N = 97 to 126) |
|---|---|---|---|---|---|
| Assembling Objects | 40 | 16 | .79 | .92 | .74 |
| Object Rotation | 90 | 7.5 | .86 | .97 | .75 |
| Mazes | 24 | 5.5 | .78 | .89 | .71 |
| Path | 44 | 8 | .82 | .92 | .64 |
| Shapes | 54 | 16 | .82 | .92 | .70 |
| Reasoning 1 | 30 | 12 | .78 | .83 | .64 |
| Reasoning 2 | 32 | 10 | .63 | .65 | .57 |
| Orientation 1 | 150 | 10 | .92 | .98 | .67 |
| Orientation 2 | 24 | 10 | .89 | .88 | .80 |
| Orientation 3 | 20 | 12 | .88 | .90 | .84 |

once again, the Reasoning 2 test yields the lowest value of all.

These data, along with information about psychometric test characteristics and test intercorrelations, were used to identify the tests for inclusion in the Trial Battery administered Summer and Fall, 1985. Paper-and-pencil test intercorrelations are examined following the discussion of the computerized tests. At that time we also describe the paper-and-pencil measures selected for the Trial Battery. Now we turn to a description of the activites involved in developing the cognitive measures included in the computer test battery.

## Computerized Cognitve/Perceptual Tests

Traditional paper-and-pencil measures such as those described above and those contained in the ASVAB allow us to assess accuracy in test performance. Advances in microcomputer development, however, permit us to examine another area of test performance, speed of response. Therefore, to identify cognitive/perceptual ability measures for inclusion in the computer battery, several sources were examined.

First, we examined the more recent theories of cognitive abilities assessment. For example, Keyes (1985) conducted an indepth review of the reaction time construct. Results from this review guided our thinking about new measures to supplement information obtained on the ASVAB. In addition, results from the review suggested several variables or parameters of interest for test development and test scoring purposes.

Second, several members of our staff conducted site visits of facilities such as the Air Force Human Research Laboratory, or AFHRL, in which extensive research in computerized testing is currently being conducted. Information gleaned from these sources also guided our test development plans.

Third, examination of highly speeded, traditional paper-and-pencil measures helped to generate some basic research ideas. For example, in a review of the psychomotor literature, McHenry (1985) reports that scores on traditional measures of perceptual speed and accuracy correlate moderately with scores on measures of wrist-finger speed or answer sheet marking. (The median value across seventeen coefficients was .38.) This suggests that for some cognitive ability constructs (or perhaps all), scores on traditional paper-and-pencil measures capture only one aspect of the response, namely accuracy. Including a speed of response component may provide information about other abilities.

A final source used to generate computer test ideas stems from the research conducted during World War II in the Aviation Psychology Program. During this research program, researchers experimented with measures of new constructs, such as Movement Judgment (Gibson, 1947). Apparatus available at that time resulted in lower than desirable levels of reliability of measurement of such constructs. With advances in computer development and computerized testing, however, these constructs

may be more reliably measured and may provide information not currently supplied by the ASVAB.

Below we describe the cognitive ability constructs measured in the computer battery and identify the tests designed to measure each. The tests themselves are described on Table 4. Following this is a discussion of the issues related to developing and scoring computerized tests.

Reaction Time/Processing Efficiency

This involves speed of reaction to stimuli or the speed with which a person perceives the stimulus independent of any time taken by the motor response component of the classic reaction time measure. The basic paradigm for this task stems from Jensen's research in which he designed a procedure to obtain independent measures of decision time and movement time. Decision time refers to the time that elapses between stimulus presentation and initiation of the response. Movement time refers to the time that elapses between initiation of a response (movement from a standard or "home" position) and the actual response (Jensen, 1982). All tests included in the computer battery that involve a reaction time component were designed to capture the two components of total reaction time.

Two measures were constructed to assess Processing Efficiency. These include Simple Reaction Time and Choice Reaction Time. In both measures, subjects respond to very simple or non-complex stimuli such as the word YELLOW or BLUE.

Memory

This involves the rate at which one observes, searches, and recalls information. Two measures were designed to assess memory. In both measures, the stimuli and task required are more complex than those included in the Processing Efficiency tests. Each test is described briefly below.

Short Term Memory: The model for this test is a memory search task introduced by S. Sternberg (1966, 1969). In the test designed for the computer battery, the subject is presented with a set of items. This disappears from the screen and the subject is then presented with a probe. The task is to indicate as rapidly as possible whether or not the probe appeared in the first set of items. Items presented to each subject consist of familiar objects (letters) or novel objects (symbols).

Number Memory: This test was modeled after a number memory test developed by Dr. Raymond Christal at AFHRL (1983). In this test, subjects are presented with simple arithmetic problems (i.e., addition, subtraction, multiplication, and division). The task differs from the traditional number operations test in that the subject is presented with one part of the problem at a time and one item may contain two, three, or more parts. After the subject reviews one part of the problem, s/he must press a button to go on to another part of the problem. At the end of the problem, the subject is presented with a possible solution and

must indicate whether it is true or false. The task here is for the subject to perform simple arithmetic operations and to recall information from previous parts of the problem.

Perceptual Speed and Accuracy

This involves the ability to perceive visual information quickly and accurately and to perform processing tasks with stimuli (e.g., make comparisons). Two measures were developed to assess this construct. Each is described briefly below.

Perceptual Speed and Accuracy: This task requires that the subject compare two sets of verbal information (e.g., numbers, letters, symbols). As indicated earlier in this paper, the ASVAB currently contains a measure of perceptual speed and accuracy. This measure was included in the computer test battery to address a basic research question. That is, does a computerized version of a highly speeded measure provide more information about a person's abilities with speed of response differentiated from accuracy?

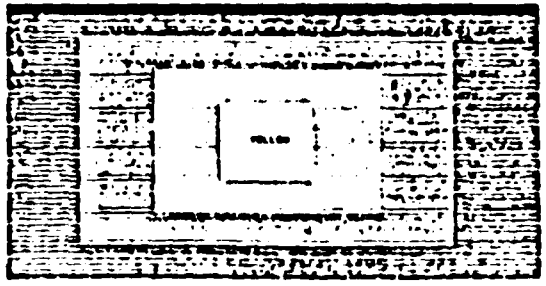
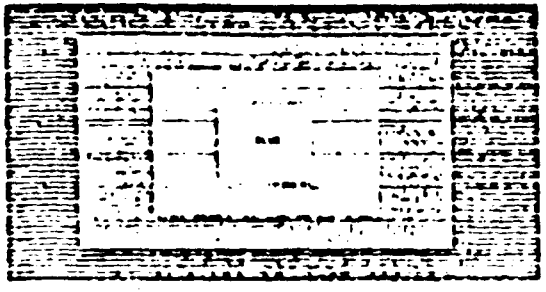
Target Identification: This test asks subjects to compare several different figures to identify the one figure that matches a target figure. This test was constructed to assess an ability required in all MOS, but particularly in combat MOS--rapidly identifying enemy equipment and vehicles.

Movement Judgment

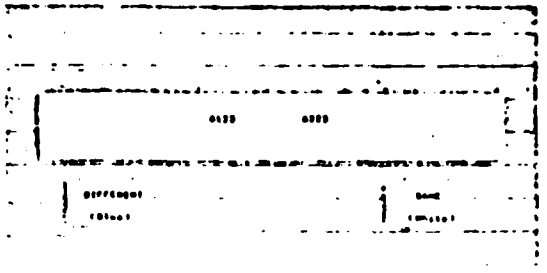
This involves the ability to judge the relative speed and direction of one or more moving objects in order to determine

Table 4

Description of Cognitive/Perceptual Computer Tests

| CONSTRUCT/MEASURE | DESCRIPTION OF TEST | SAMPLE ITEM |
|---|---|---|
| **REACTION TIME** | | |
| Simple Reaction Time | The subject is instructed to place his/her hands on the green "home" buttons or in the Ready position. When the subject's hands are in the Ready position, a small box appears on the screen. After a delay period which varies from 1.5 to 3.0 seconds, the word YELLOW appears in the box. At this point, the subject must remove his/her preferred hand from the "home" buttons to strike the Yellow key on the testing panel. The subject must then return his/her hands to the ready position to receive the next item. The test contains 15 items. Although it is self-paced, subjects are given 10 seconds to respond before the computer times out and prepares to present the next item. |  |
| Choice Reaction Time | Choice reaction time is assessed for two response alternatives only. This measure is obtained in virtually the same manner as the simple reaction time measure. The major difference involves stimulus presentation. Rather than presenting the same stimulus (YELLOW) on each trial, the stimulus varies. That is, subjects may see either of the stimuli BLUE or WHITE on the computer screen. When the stimulus appears, the subject is instructed to move his/her preferred hand from the "home" keys to strike the key that corresponds with the term (BLUE or WHITE) appearing on the screen. This test contains 15 items. Although the test is self-paced, the computer is programmed to allow the subject nine seconds to respond before going on to the next item. |  |

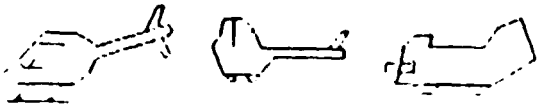
/continued

| CONSTRUCT/MEASURE | DESCRIPTION OF TEST | SAMPLE ITEM |
|---|---|---|
| PERCEPTUAL SPEED AND ACCURACY | | |
| Perceptual Speed and Accuracy | This test is designed to measure the ability to compare rapidly two visual stimuli presented simultaneously and determine whether they are the same or different. At the beginning of each trial, the subject is instructed to hold down the home keys. After a brief delay, the stimuli are presented. The subject must decide next whether the stimuli are the same or different. He/she must then depress a white button if the stimuli are the same or a blue button if the stimuli are different. Four different "types" of stimuli are used: alpha, numeric, symbolic, and words. Within the alpha, numeric, and symbolic stimuli, the length of the stimulus is varied. Three different levels of length are presented: two-character, five-character, and nine-character. The test consists of 48 trials. The primary dependent variable is the subject's average response time across all trials in which the subject makes a correct response. |  |



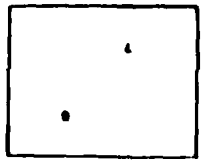
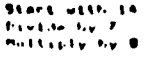
/continued

Cognitive/Perceptual Computer Tests

| CONSTRUCT/MEASURE | DESCRIPTION OF TEST | SAMPLE ITEM |
|---|---|---|
| PERCEPTUAL SPEED AND ACCURACY (continued)<br><br>Target Identification | This test was designed to be a job-relevant measure of perceptual speed and accuracy. In this test, the subject is presented with a target object and three stimulus objects. The objects are pictures of military vehicles or aircraft (e.g., tanks, planes, helicopters). The target object is the same as one of the stimulus objects. However, the target may be rotated or reduced in size relative to its stimulus counterpart, or the target may be "moving" and growing across the screen. The subject must determine which of the three stimulus objects is the same as the target object and then press a button on the response pedestal corresponding to that choice. The test consists of 48 items; 24 are stationary, 24 are moving. The primary dependent variable is the subject's average response time across all trials in which the subject makes a correct response. | <br>TARGET<br>  <br>BLUE YELLOW WHITE |

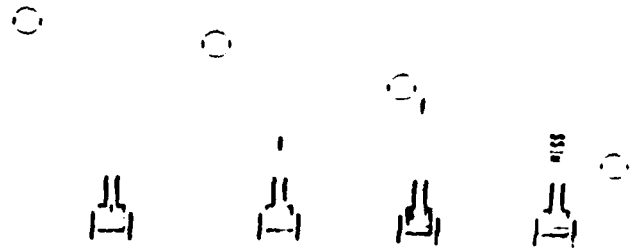


/continued

| CONSTRUCT/MEASURE | DESCRIPTION OF TEST | SAMPLE ITEM |
|---|---|---|
| **MEMORY** | | |
| Short Term Memory | At the computer console, the subject is instructed to place his/her hands on the green home buttons. The first stimulus set then appears on the screen. A stimulus contains one, three, or five objects (letters or symbols). Following a delay period, the stimulus set disappears. When the probe appears, the subject must decide whether or not it was part of the stimulus set. If the probe was present in the stimulus set, the subject must strike the white key. If the probe was not present, the subject must strike the blue key on the response pedestal. The test includes 48 items. The primary dependent variable is the subject's average response time across those trials in which the subject makes a correct response. |  Item set Probe |
| Number Memory | At the beginning of each trial of this test, the subject is presented with a single number on the computer screen. After studying the number, the subject is instructed to push a button to receive the next part of the problem. When the subject presses the button, the first part of the problem disappears and another number appears along with an operation term (e.g., "Add 9" or Subtract 6"). Once the subject has combined the first number with the second, he/she must press a button to receive a new number and operation term. This procedure continues until a solution to the problem is presented. The subject must then indicate whether the solution presented is correct or incorrect. In total, the test consists of 27 such items. |  |



/continued

| CONSTRUCT/MEASURE | DESCRIPTION OF TEST | SAMPLE ITEM |
|---|---|---|
| MOVEMENT JUDGMENT<br><br>Cannon Shoot | At the beginning of each trial of this test, a stationary cannon appears on the computer console. The starting position of this cannon varies from trial to trial (i.e., it is positioned on the top, bottom, or side of the screen). The cannon is capable of firing a shell. The shell travels at a constant speed on each trial. Shortly after the cannon appears, a circular target moves onto the screen. This target moves in a constant direction at a constant rate of speed throughout the trial, though the speed and direction vary from trial to trial. The subject's task is to push a response button to fire the shell such that the shell intersects the target when the target crosses the shell's line of fire. The test includes 48 items. The primary dependent variable is a deviation score indicating the difference between time of fire and optimal fire time (e.g., direct hits yield a deviation score of zero.) |  |

where those objects will be at a given point in time or when the two might intersect. Results from the Aviation Psychology Program indicate that when used to predict aircrew performance, measures of this construct yielded validities in the .20s (Gibson, 1947). Because several combat MOS require movement judgment to perform successfully on the job (e.g., Armor Crewman), the Cannon Shoot test was developed. In this test, subjects must study a moving target and determine the optimal point at which firing a shell will produce a direct hit.

### Issues Related to Test Development and Test Scoring

Issues related to scoring computerized tests impact on test development issues. For example, when computing reaction time scores, one must decide whether or not to include correct and incorrect responses. If it is appropriate to assume that reaction time scores for correct responses provide the same information as reaction time scores for incorrect responses, then item development strategies will vary from those used when the assumption is that the two will differ.

During the test development phase of the project, we began with the assumption that correct and incorrect responses provide different information. Hence, our goal in constructing items for each test was to ensure that all or nearly all subjects could, of course, respond to each item as well as identify the correct response. Because computer administration permits precise response time estimates, the primary emphasis for the

cognitive/perceptual tests, then, is on speed of response, rather than response accuracy.

Another concern during the test development phase involved the parameters included in each test. That is, by systematically varying test stimuli, one may uncover slightly different abilities within the same test. Thus, for each cognitive/perceptual test we relied on the literature and our knowledge of job requirements to identify parameters of interest. For example, in the Short Term Memory test, the literature suggests that familiar stimuli such as letters be used to avoid opportunity-to-learn differences among subjects. Following several site visits in which we observed soldiers performing on the job, it became clear that for several combat MOS, soldiers must very quickly and accurately discern differnces between objects such as friendly and enemy vehicles, equipment and so on. Therefore, the concept of quickly memorizing and discerning differences between novel stimuli was incorporated as another parameter in the Short Term Memory test (i.e. symbols served as the novel stimuli). Other parameters that varied in the test include number of objects appearing in the set, time allowed to observe the objects in the set, and the delay period between presentation of objects in the set and the presentation of the probe. Results from the pilot tests or tryouts were analyzed to assess the impact of the varied parameters on reaction time

scores. Then, following each tryout, we modified each computer test by refining test parameters or eliminating them if results indicated that they had little or no impact on reaction time.

In addition to assessing slightly different abilities, varied parameters also allow computation of other dependent variables in addition to reaction time. Again using the Short Term Memory test as an example, the parameter "number of objects in a set" may be used to calculate the slope and intercept. These scores are obtained by regressing mean reaction time against item set length. In terms of speed of processing, the slope represents the average increase in reaction time with an increase of one object in the stimulus set. Thus, the lower the value, the faster the access. The intercept represents all other processes not involved in memory search such as encoding the probe, determining whether or not a match has been found and executing the response.

A third issue of critical importance in computerized test development concerns the procedures used to score subjects' responses. One scoring question that we addressed early in the test development stage involves whether or not reaction time scores from incorrect responses are to be included in total reaction time scores. As noted above, we decided to include reaction time scores from correct responses only.

Another scoring question concerns the method of computing reaction time scores. For example, in our description of the

Processing Efficiency construct, we noted that Jensen (1982) views total reaction time as comprising two distinct components--Decision Time and Movement Time. Following the first pilot test, for each test we computed both Decision and Movement time scores for each test as well as Total time. Analyses indicated that Decision and Total time are very highly correlated, so we elected to use Total time as the dependent variable for each test.

Following on the heels of this issue is the question about how to estimate an individual's reaction time score in the best way. In other words, is the mean preferred over the median? Analysis of pilot test results suggest that the mean is slightly more reliable than the median (i.e. test-retest reliability estimates). Therefore, for all reaction time measures, scores are computed using the mean value. The one exception to this involves the scores computed for Simple and Choice Reaction Time. Because these tests contained only 15 items, a single extreme response could result in an unrepresentative mean score. Therefore, a trimmed mean score was computed. This score consists of the mean computed over all responses with the highest and lowest values omitted.

A final scoring issue concerns missing data. In other words, given that a single subject may not obtain a perfect score on a particular test, some information is missing when computing

the mean total reaction time, slope and intercept for that subject. Thus, we established a maximum number of missing items permitted for each test. This limit for all tests, with the exception of Number Memory, is set at ten percent. Hence, for Simple and Choice Reaction Time subjects may miss up to two items, for Short Term Memory, Perceotual Speed and Accuracy, and Target Identification the value is set at five. Because Number Memory requires subjects to provide several responses for a single item, the possibility of missing data is higher. To ensure that sufficient numbers of subjects were available for analysis, we permitted subjects to miss up to seven of the twenty-seven items. (It is important to note that the test itself was modified to reduce the likelihood of missing data for subsequent administrations. These modifications are described later in this paper.)

Throughout this discussion of issues in scoring, we have not yet touched upon the Cannon Shoot test. That is because procedures for scoring this test differed from those used to score the other cognitive/perceptual tests. A reaction time score for this test is inappropriate because the task requires the subject to ascertain the optimal time to fire to ensure a direct hit on the target. Therefore, responses on this measure were scored by computing a deviation score that is composed of the difference between the time the subject fired and the optimal time. These scores are summed across all items for each subject

and a mean deviation time score is computed.

Evaluating the Computerized Cognitive/Perceptual Tests

To evaluate the computerized tests, the battery was administered at three tryout sites. These include Fort Carson, Fort Lewis, and Fort Knox. When administered at Fort Carson, the computer battery contained only a few of the tests described above. Thus, the bulk of the information used to evaluate the tests came from the data obtained at the Fort Lewis and Fort Knox tryouts. In general, we considered three aspects when determining how to improve or modify each test. These include test content factors, stimulus presentation and response format factors, and reliability evidence. Each is described briefly below.

The test content factors include the parameters used to vary stimulus presentation. For each test, then, we examined the effects of different parameters and made decisions about modifying test content. For example, in the Short Term Memory Test, we examined two levels of item set display periods (.5 secs. vs. 1.0 sec.) and two levels of probe delay period (2.5 secs. vs. 3.0 secs.). Results from the first tryout revealed that neither the display period nor the delay period had any impact on reaction time. Therefore, for the next tryout we decided to experiment with other levels of these parameters (e.g., the probe delay period contained two levels, .5 sec. and 2.5 secs.).

Another factor considered in modifying the test involves the stimulus presentation and response format. This involved more of an "armchair analysis" to isolate changes required on each test. For example, we compared the different sets of test instructions to identify those that appeared to be working well and those that required modifications. For these changes we also relied on interviews with subjects to identify test instructions that seemed unclear or difficult to understand.

Also at this time, we compared the response formats across tests. In other words, for each test the computer program was designed to highlight the response alternatives and then, after the subject responds, provide feedback on the selected response. Our discussion with subjects participating in the first tryout indicated that some formats were clearly better than others. Based on this information, then, the program and test files were modified to reflect the desired changes.

A third and final factor considered in the test modification activites was reliability of the measures. Internal consistency reliability estimates were computed for all dependent measures following each tryout. Results from the Fort Knox tryout, which contains the largest sample tested, are provided in Table 5. On this table, we have listed the fifteen dependent measures for the seven cognitive/perceptual ability tests developed. The table contains means, standard deviations, internal consistency reliability estimates computed on a sample of 256 and test-retest

Table 5

Characteristics of the 19 Computer Test Dependent Measures
Ft. Knox (N=256)[a]

| Dependent Measure | Mean | SD | Reliability $r_{sh}$ | Reliability $r_{tt}$[b] |
|---|---|---|---|---|
| COGNITIVE/PERCEPTUAL | | | | |
| Simple Reaction Time - Mean RT | 56.23 hs | 18.83 hs | .90 | .37 |
| Choice Reaction Time - Mean RT | 67.41 hs | 10.20 hs | .89 | .56 |
| Perc Speed & Acc - Pct Correct | 88% | 8% | .83 | .59 |
| Perc Speed & Acc - Mean RT | 325.61 hs | 70.38 hs | .96 | .65 |
| Perc Speed & Acc - Slope | 42.74 hs/ch | 15.56 hs/ch | .88 | .67 |
| Perc Speed & Acc - Intercept | 67.96 hs | 45.02 hs | .74 | .55 |
| Target ID - Pct Correct | 90% | 10% | .84 | .19 |
| Target ID - Mean RT | 528.70 hs | 133.96 hs | .96 | .67 |
| Short Term Mem - Pct Correct | 85% | 8% | .72 | .34 |
| Short Term Mem - Mean RT | 129.68 hs | 23.84 hs | .94 | .78 |
| Short Term Mem - Slope | 7.22 hs/ch | 4.53 hs/ch | .52 | .47 |
| Short Term Mem - Intercept | 108.12 hs | 23.18 hs | .84 | .74 |
| Cannon Shoot - Time Score | 78.60 hs | 20.28 hs | .88 | .66 |
| Number Memory - Pct Correct | 83% | 13% | .63 | .53 |
| Number Memory - Mean Oper Time | 230.71 hs | 73.92 hs | .95 | .88 |
| PSYCHOMOTOR | | | | |
| Target Track 1 - Mean Log Dist | 3.22 | .44 | .97 | .68 |
| Target Track 2 - Mean Log Dist | 3.91 | .49 | .97 | .77 |
| Targ Sht - Mean Fire Time (std) | -.01 | .48 | .91 | .48 |
| Targ Sht - Mean Log Dist (std) | -.01 | .41 | .86 | .58 |

a. N=256, but varies slightly from test to test.
b. N=120 for test-retest reliabilities, but varies slightly from test to test.

reliability estimates computed on a sample of about 120.

Note that for the internal consistency estimates the highest values appear for mean reaction time scores for Short Term Memory, Number Memory, Peceptual Speed and Accuracy, and Target Identification. The lowest values appear for Short Term Memory Slope and Number Memory Percent Correct. These values, with the exception of the two above, appear comparable to the estimates for the paper-and-pencil measures.

The lowest values appear for Percent Correct values for Target ID and Short Term Memory as well as for Simple Reaction Time-Mean RT. The highest values appear for Number Memory, Short Term Memory-Mean RT and Short Term Memory Intercept. Overall, we were fairly impressed with the test-retest results. With the exception of a few low values, these estimates appear to be similar to those reported for the paper-and-pencil measures.

The remaining reliability estimates included on the table--those reported for the Psychomotor tests--are discussed in detail by McHenry & McGue (1985). Following the discussion of the Psychomotor measures, we examine the intercorrelations among the new paper-and-pencil measures, all computer tests, and subtests of the ASVAB. Results from a factor analysis of scores on all the measures listed above will also be reported. Finally, we describe the modifications made to the paper-and-pencil measures and computer tests made after the Fort Knox tryout. The battery resulting after these modifications were made is referred to as

the Trial Battery.

### Intercorrelations and Factor Analysis Results

Table 6 contains the intercorrelations for the ASVAB subtests, paper-and-pencil cognitive measures, and the computer tests which include both cognitive/perceptual and psychomotor measures. Note that we have also included scores on the AFQT. Before looking at the correlations, it is important to highlight two facts. First, the sample used to generate these intercorrelations includes only those subjects with test scores available on all variables (N=168). Second, these data represent the relationships among the tests administered at Fort Knox. Because of changes made on the tests following the Fort Knox tryout, we can expect some changes in these data.

In examining these relationships, first look at the correlations between tests within the same battery. For example, correlations between ASVAB subtests range from .02 to .74 (absolute values). The range of intercorrelations is more restricted when examining the relationships between the cognitive paper-and-pencil tests (.27 to .67). This range of values reflects the fact that these measures were designed to tap very similar cognitive constructs. Intercorrelations for the cognitive/perceptual computer tests range from .00 to .83 in absolute terms. Note that the highest values appear for correlations between variables computed from the same test. For example, the correlation between Short Term Memory Reaction Time

Table 6

Intercorrelations among the ASVAB and Pilot Trial Battery (PTB) Tests
Ft. Knox Sample (N=168)

| | | AFQT Score | Gen Science | Arith Reas | Word Know | Parg Comp | Numb Ops | Code Spd | Auto/Shop | Math Know | Mech Comp | Elec Info | Assmbl Obj | Obj Rotat | Shapes | Maze | Path | Reas 1 | Reas 2 | Orient 1 | Orient 2 | Orient 3 | SRT-RT | CRT-RT | PSEA-PC | PSEA-RT | PSEA-Slp | PSEA-Int | Targ ID-PC | Targ ID-RT | STM-PC | STM-RT | STM-Slp | STM-Int | Con Shoot | No Rem-PC | No Rem-RT | Trk 1-Dist | Trk 2 Dist | TSht Time | TSht Dist |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASVAB | AFQT Score | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Gen Scienc | 61 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Arith Reas | 87 | 54 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Word Know | 81 | 66 | 61 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Parg Comp | 69 | 43 | 53 | 58 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Numb Ops | 44 | 02 | 21 | 06 | 14 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Code Spd | 30 | -02 | 14 | 10 | 14 | 55 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Auto/Shop | 45 | 54 | 50 | 45 | 29 | -08 | -07 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Math Know | 76 | 60 | 74 | 62 | 54 | 19 | 15 | 43 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Mech Comp | 55 | 50 | 62 | 54 | 36 | -10 | -06 | 44 | 57 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Elec Info | 56 | 66 | 56 | 55 | 39 | -03 | 01 | 71 | 59 | 63 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PTB - Paper-and-Pencil Tests | Assmbl Obj | 44 | 38 | 48 | 40 | 29 | -01 | 19 | 39 | 40 | 57 | 58 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Obj Rotat | 38 | 20 | 33 | 18 | 18 | 18 | 13 | 27 | 29 | 36 | 32 | 47 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Shapes | 35 | 29 | 33 | 28 | 20 | 14 | 23 | 17 | 36 | 35 | 26 | 51 | 50 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Maze | 32 | 23 | 38 | 16 | 12 | 19 | 19 | 34 | 36 | 44 | 35 | 60 | 58 | 67 | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Path | 35 | 20 | 37 | 21 | 17 | 24 | 31 | 33 | 35 | 42 | 32 | 54 | 59 | 58 | 63 | | | | | | | | | | | | | | | | | | | | | | | | |
| | Reas 1 | 46 | 33 | 48 | 41 | 33 | 06 | 15 | 29 | 47 | 51 | 32 | 66 | 38 | 41 | 51 | 52 | | | | | | | | | | | | | | | | | | | | | | | |
| | Reas 2 | 41 | 32 | 44 | 34 | 34 | 02 | 12 | 18 | 49 | 38 | 30 | 52 | 27 | 44 | 33 | 37 | 50 | | | | | | | | | | | | | | | | | | | | | | |
| | Orient 1 | 45 | 41 | 45 | 35 | 29 | 14 | 20 | 40 | 47 | 50 | 39 | 52 | 55 | 54 | 55 | 58 | 49 | 61 | | | | | | | | | | | | | | | | | | | | | |
| | Orient 2 | 50 | 29 | 52 | 38 | 33 | 18 | 15 | 36 | 48 | 49 | 35 | 51 | 38 | 42 | 49 | 42 | 48 | 47 | 55 | | | | | | | | | | | | | | | | | | | | |
| | Orient 3 | 59 | 51 | 62 | 48 | 43 | 05 | 15 | 55 | 60 | 63 | 56 | 61 | 42 | 42 | 51 | 46 | 49 | 53 | 67 | 56 | | | | | | | | | | | | | | | | | | | |
| PTB - Computerized Cognitive-Perceptual Tests | SRT-RT | -23 | -11 | -20 | -15 | -15 | -25 | -07 | -09 | -14 | -11 | -08 | -11 | -21 | -09 | -22 | -23 | -29 | -14 | -17 | -31 | -18 | | | | | | | | | | | | | | | | | | |
| | CRT-RT | -16 | -05 | -21 | -02 | -08 | -20 | -16 | -18 | -10 | -14 | -16 | -16 | -37 | -19 | -37 | -33 | -21 | -12 | -27 | -37 | -22 | 53 | | | | | | | | | | | | | | | | | |
| | PSEA-PC | 25 | -01 | 22 | 18 | 23 | 14 | 30 | 04 | 12 | 11 | 02 | 18 | 14 | 10 | 02 | 19 | 25 | 14 | 13 | 10 | 14 | 17 | 17 | | | | | | | | | | | | | | | | |
| | PSEA-RT | -06 | -04 | -01 | -02 | 00 | -14 | -15 | -09 | -06 | -12 | -10 | -15 | -17 | -24 | -36 | -18 | -03 | 01 | -13 | -16 | -13 | 19 | 31 | 50 | | | | | | | | | | | | | | | |
| | PSEA-Slp | 11 | -00 | 17 | 07 | 09 | -02 | -01 | -05 | 06 | 01 | -14 | 04 | -07 | -12 | -15 | -01 | 12 | 14 | 00 | 01 | -01 | -03 | 09 | 52 | 82 | | | | | | | | | | | | | | |
| | PSEA-Int | -23 | -03 | -30 | -10 | -12 | -13 | -15 | -04 | -14 | -17 | -01 | -31 | -21 | -16 | -28 | -28 | -30 | -25 | -20 | -26 | -15 | 32 | 31 | -27 | -08 | -61 | | | | | | | | | | | | | |
| | Targ ID-PC | 12 | 07 | 07 | 21 | 09 | -08 | 10 | 07 | 06 | 12 | 01 | 12 | 05 | 09 | -05 | 13 | 12 | 14 | 05 | 03 | 11 | 11 | 07 | 40 | 33 | 32 | -13 | | | | | | | | | | | | |
| | Targ ID-RT | -23 | -31 | -29 | -19 | -06 | -05 | -02 | -34 | -25 | -37 | -32 | -41 | -46 | -34 | -50 | -45 | -29 | -20 | -36 | -37 | -35 | 23 | 42 | 20 | 47 | 32 | 13 | 16 | | | | | | | | | | | |
| | STM PC | 21 | 05 | 16 | 14 | 12 | 17 | 25 | 01 | 05 | 12 | 05 | 15 | 10 | 15 | 02 | 10 | 14 | 23 | 10 | 20 | 20 | -06 | 04 | 50 | 17 | 26 | -25 | 25 | 05 | | | | | | | | | | |
| | STM RT | -02 | 02 | -10 | 11 | 05 | -09 | -13 | -02 | -01 | -09 | -08 | -13 | -20 | -13 | -20 | -22 | -06 | -05 | -12 | -20 | -06 | 23 | 40 | 26 | 49 | 25 | 23 | 27 | 47 | 08 | | | | | | | | | |
| | STM-Slp | 12 | 06 | 09 | 07 | 10 | 10 | 06 | 01 | 08 | 11 | -02 | 09 | -00 | 05 | 02 | 10 | 10 | 09 | 05 | 06 | 11 | -06 | 03 | 28 | 29 | 26 | -11 | 18 | 13 | 32 | 39 | | | | | | | | |
| | STM-Int | -10 | -02 | -17 | 08 | -00 | -16 | -17 | -03 | -06 | -16 | -08 | -19 | -22 | -16 | -31 | -29 | -12 | -11 | -17 | -25 | -14 | 28 | 42 | 10 | 35 | 11 | 31 | 18 | 42 | -11 | 03 | -19 | | | | | | | |
| | Con Shoot | -11 | -13 | -14 | -13 | 01 | -04 | -03 | -17 | -03 | -15 | -15 | -14 | -17 | -07 | -15 | -17 | -08 | -03 | -11 | -08 | -08 | 13 | 10 | 00 | 08 | -01 | 11 | -09 | 25 | -02 | 25 | 14 | 10 | | | | | | |
| | No Rem-PC | 60 | 42 | 58 | 41 | 36 | 35 | 20 | 33 | 56 | 35 | 34 | 35 | 26 | 25 | 29 | 32 | 40 | 31 | 37 | 33 | 43 | -16 | -09 | 29 | 02 | 15 | -20 | 14 | -12 | 23 | -00 | 02 | -01 | -18 | | | | | |
| | No Rem-RT | -49 | -27 | -46 | -22 | -21 | -45 | -29 | -23 | -42 | -23 | -31 | -17 | -26 | -25 | -34 | -26 | -20 | -18 | -35 | -29 | -36 | 21 | 24 | 11 | 34 | 21 | 03 | 10 | 27 | 07 | 18 | 15 | 11 | 08 | -45 | | | | |
| PTB - Computerized Psychomotor Tests | Trk 1-Dist | -20 | -21 | -31 | -16 | -07 | 09 | -06 | -33 | -24 | -44 | -28 | -38 | -34 | -23 | -35 | -41 | -29 | -26 | -28 | -39 | -36 | 14 | 25 | -12 | 08 | 00 | 11 | -04 | 42 | -29 | 25 | -15 | 35 | 27 | -14 | -00 | | | |
| | Trk 2-Dist | -19 | -23 | -28 | -15 | -03 | 07 | -05 | -33 | -21 | -38 | -29 | -37 | -28 | -14 | -32 | -35 | -22 | -22 | -25 | -36 | -32 | 11 | 19 | -01 | 11 | 04 | 09 | 02 | 39 | -19 | 25 | -01 | 27 | 30 | -14 | 02 | 81 | | |
| | TShT Time | -09 | -08 | -16 | -03 | 05 | -02 | -00 | -18 | -10 | -21 | -17 | -10 | -14 | -02 | -23 | -17 | -11 | -01 | -09 | -14 | -06 | 08 | 16 | 16 | 22 | 09 | 12 | 12 | 32 | 09 | 22 | 15 | 15 | 12 | -10 | 15 | 23 | 19 | |
| | TShT Dist | -09 | -19 | -17 | -09 | -02 | 13 | 00 | -26 | -12 | -31 | -23 | -31 | -22 | -14 | -30 | -37 | -24 | -14 | -23 | -24 | -28 | 08 | 16 | -07 | 03 | -00 | 09 | -11 | 32 | -12 | 27 | -16 | 38 | 25 | -08 | 02 | 60 | 55 | -15 |

and Intercept is .83, while the correlation between Perceptual Speed and Accuracy Slope and Reaction Time is .82. Intercorrelations between psychomotor variables range from .15 to .81 in absolute terms. Note that scores on the two tracking tests correlate the highest.

Perhaps the most important question to consider is the similarity between the different groups of measures. In other words, do the paper-and-pencil measures and computer tests correlate highly with the ASVAB or are they actually measuring unique or different abilities? To address this question, in part, examine the intercorrelations between the ASVAB and other groups of tests. First, for the paper-and-pencil tests, these correlations range from .01 (Assembling Objects and Number Operations) to .63 (Orientation 3 and Mechanical Comprehension). Note that across all paper-and-pencil tests, ASVAB Mechanical Comprehension appears to correlate the highest with the new tests. Across all ASVAB subtests, Orientation 3 yields the highest correlations.

Now consider the correlations between the ASVAB subtests and the computerized cognitive/perceptual tests. In absolute terms, these values range from .00 (Paragraph Comprehension and Perceptual Speed and Accuracy, or PS&A Reaction Time and Short Term Memory Intercept) to .58 (Arithmetic Reasoning and Number Memory Percent Correct). Across all ASVAB subtests, scores on the Short Term Memory Reaction Time and Slope yield the lowest

correlations. The highest values appear for Number Memory Percent Correct and Reaction Time.

Correlations between ASVAB subtests and psychomotor variables range from .00 (Target Shoot-Time and Coding Speed and Target Shoot-Distance and Coding Speed) to -.44 (Mechanical Comprehension and Tracking 1). Note that for the most part, these four variables yield the highest correlations with Mechanical Comprehension and Electronics Information. The lowest correlations appear for Paragraph Comprehension, Number Operations and Coding Speed.

Briefly, the intercorrelations between the paper-and-pencil tests and the computerized tests in general range from .00 to .46 (in absolute terms). The computerized test variables that correlate consistently highly with the paper-and-pencil tests include Target ID-Reaction Time, Number Memory Percent Correct and Reaction Time, Tracking 1, and Tracking 2. Intercorrelations between the cognitive/perceptual computer tests and the psychomotor computer tests range from .00 to .42. The highest values appear for the correlations between the four psychomotor measures and Target ID Percent Correct and Short Term Memory Slope.

In addition to examining the intercorrelations among all the cognitive/perceptual measures and psychomotor measures, we also examined results from a factor analysis. It is important to note here that two variables, PS&A Reaction Time and Short Term Memory

Reaction Time, were omitted from this analysis. This is because the reaction time scores from these measures correlated very highly with their corresponding Slope or Intercept variables. To avoid obtaining communalities greater than one, these two reaction time measures were omitted. Results from the seven factor solution of a principal components factor analysis with varimax rotation are displayed in Table 7.

Our interpretation of these data are described, by factor, below.

Factor 1 includes eight of the ASVAB subtests (GS, AR, WK, PC, AS, MK, MC, and EI), six of the paper-and-pencil measures (*Assembling Objects*, *Reasoning* 1 and 2, and Orientation 1, 2, and 3) and two cognitive/perceptual computer variables (Number Memory Percent Correct and Reaction Time). Because this factor contains measures of verbal, numerical and reasoning ability we have termed this "g".

Factor 2 includes all of the cognitive paper-and-pencil measures, Mechanical Comprehension from the ASVAB, and Target ID Reaction Time from the computer tests. This is, then, a general Spatial factor.

Table 7 Results from a Principal Components Factor Analysis of Scores on the ASVAB, Cognitive Paper-and-Pencil Measures, and Cognitive/Perceptual and Psychomotor Computer Tests[a]
(N = 168)

| Variable | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor 6 | Factor 7 | $h^2$ |
|---|---|---|---|---|---|---|---|---|
| ASVAB GS | 75 | | | | | | | 59 |
| ASVAB AR | 75 | | | | | | | 73 |
| ASVAB WK | 77 | | | | | | | 62 |
| ASVAB PC | 62 | | | | | | | 47 |
| ASVAB NO | | | | | | 84 | | 77 |
| ASVAB AS | 62 | | | | | 62 | | 58 |
| ASVAB MK | 77 | | | | | | | 70 |
| ASVAB MC | 63 | 38 | -30 | | | | | 68 |
| ASVAB EI | 72 | | | | | | | 65 |
| Assemb Obj | 35 | 69 | | | | | | 66 |
| Obj Rotation | | 61 | | | | | | 49 |
| Shapes | | 66 | | | | | | 51 |
| Mazes | | 70 | | | | | | 67 |
| Path | | 67 | -30 | | | | | 65 |
| Reason 1 | 37 | 58 | | | | | | 54 |
| Reason 2 | 37 | 47 | | | | | | 44 |
| Orient 1 | 37 | 64 | | | | | | 58 |
| Orient 2 | 40 | 46 | | | -30 | | | 52 |
| Orient 3 | 60 | 52 | | | | | | 67 |
| SRT-RT | | | | | 63 | | | 44 |
| CRT-RT | | | | | 61 | | | 50 |
| PS&A-PC | | | | 67 | 31 | | | 70 |
| PS&A Slope | | | | 88 | | | | 81 |
| PS&A Inter | | | | -65 | 50 | | | 74 |
| Target ID-PC | | | | 40 | | | | 25 |
| Target ID-RT | | -41 | 37 | | 30 | | | 57 |
| STM-PC | | | | 39 | | | 34 | 41 |
| STM-Slope | | | | | | | 41 | 25 |
| STM-Int | | | 38 | | 51 | | | 47 |
| Cannon Shoot | | | 32 | | | | | 19 |
| NM-PC | 53 | | | | | 37 | | 52 |
| NM-RT | -37 | | | | | -46 | | 54 |
| Tracking 1 | | | 86 | | | | | 82 |
| Tracking 2 | | | 77 | | | | | 66 |
| Target Shoot-TF | | | | | | | 42 | 23 |
| Target Shoot-Dist | | | 64 | | | | | 48 |
| Variance Explained | 5.69 | 4.70 | 2.83 | 2.37 | 1.92 | 1.87 | 1.17 | |

[a]Note that the following variables were not included in this factor analysis: APQT, PS&A Reaction Time and Short Term Memory Reaction Time.

(Please also note that decimals have been omitted.)

Factor 3 comprises the three psychomotor tests (Tracking 1, Tracking 2, and Target Shoot Distance), three cognitive/perceptual computer test variables (Target ID Reaction Time, Short Term Memory Intercept, and Cannon Shoot), the Path Test, and Mechanical Comprehension from the ASVAB. Given the high loadings of the psychomotor tests on this factor, we refer to this as the Motor factor.

Factor 4 includes variables from the cognitive/perceptual computer tests. These include PS&A Percent Correct, Slope and Intercept, Target ID Percent Correct, and Short Term Memory Percent Correct. This factor appears to involve Accuracy.

Factor 5 again contains variables from the cognitive/perceptual tests including Simple Reaction Time RT, Choice Reaction Time RT, Short Term Memory Intercept, PS&A Intercept and Percent Correct, and Target ID RT. Also loading on this factor is a paper-and-pencil test Orientation 2. We refer to this factor as the Speed factor.

Factor 6 contains four variables, two from the ASVAB (Number Operations and Coding Speed) and two from the cognitive/perceptual computer tests (Number Memory Percent Correct and Reaction Time). This factor appears to

represent both Speed and Number Ability.

Factor 7 contains three variables from the computer tests. These include Short Term Memory Percent Correct and Slope, and Target Shoot Time to Fire. Although this factor is more difficult to interpret, we believe that it represents a response style factor. That is, this factor suggests that those individuals who take a longer time to fire on the Target Shoot Test also tend to have higher slopes on the Short Term Memory (lower processing speeds with increased bits of information) but are more accurate or obtain higher percent correct values on the Short Term Memory test.

## Modifications for the Trial Battery

To prepare for large scale administration of the battery of predictors, several modifications were warranted. This is primarily because of the reduced time allotted for predictors during the Summer and Fall 1985 Concurrent Validation study. Decisions about how to pare down the battery are reported separately for paper-and-pencil measures and computer measures.

### Paper-and-Pencil Measures

Factors involved in this decision process include (1) time available; (2) the goal of ensuring broad coverage of cognitive constructs; (3) reliability data; and (4) summary validity data gleaned from the literature review. Regarding the first factor, administration time allotted for cognitive paper-and-pencil

measures was pared from a total of 167 minutes (this includes 107 minutes for actual test time and 60 minutes for test instructions) to about 100 minutes (this includes 63 minutes for actual test time and 36 minutes for test instructions). Thus, about one-third of the tests were to be dropped.

As noted above, one of the major goals was to ensure broad coverage of the cognitive ability domain. Hence, we first considered retaining at least one measure from each of the five construct areas.

As a first cut, the decision about whether to retain or drop each test was based on the construct assessed, the time required for each test, and reliability information. From this, we decided to drop Reasoning 2 (low reliability), Orientation 1 (slightly lower reliability than the other orientation measures) and the Path Test (slightly lower reliability and greater testing time than Mazes).

Tests remaining in the battery include measures of Spatial Visualization-Rotation (Assembling Objects and Object Rotation), Spatial Visualization-Scanning (Mazes), Field Independence (Shapes), Spatial Orientation (Orientation 2 and 3), and Induction (Reasoning 1). Administration time for these tests summed to 121 minutes (79 for actual test time and 42 minutes for test instructions). Thus, a second cut was in order.

To identify the test or tests for elimination, we considered

the validity evidence gleaned from the literature review as well as judgments obtained from the expert panel. Input from our Scientific Advisory Group also aided in making this final decision. The final cut, then, involved dropping the Shapes test. This is because we determined that the other measures of spatial visualization (Assembling Objects, Object Rotation, and Mazes) are more likely to be useful in predicting success in a wider variety of jobs than would a measure of field independence.

Modifications made to the paper-and-pencil measures are listed in Table 8. Note that, for the most part, all tests were retained as they appeared for the Fort Knox administration. The only exception is Assembling Objects; eight items were eliminated while the time limit of sixteen minutes was retained. Analysis of Fort Knox completion rates indicated that more information could be obtained for each subject by reducing the number of items. Item difficulty levels were examined to remove extreme items (i.e., items that yielded very high or very low difficulty levels.)

Computerized Tests

Plans for modifying the computer test battery were based on similar kinds of factors. These include time available for testing, the goal of ensuring that a broad range of constructs are assessed, and reliability of the measures. Once again, of greatest concern was the time available for the computer tests. Time estimates from the Fort Knox administration indicate that

Table 8

Changes in Paper and Pencil Measures
for the Trial Battery Cognitive Measures

| Test Name | Changes from Fort Knox to Summer 1985 Battery | Time (in minutes) |
|---|---|---|
| Assembling Objects | Decrease from 40 to 32 items | 16 |
| Object Rotation | Retain as is with 90 items | 7.5 |
| Shapes | Drop Test | 0 |
| Mazes | Retain as is with 24 items | 5.5 |
| Path Test | Drop test | 0 |
| Reasoning 1 | Retain as is with 30 items<br>New name REASONING TEST | 12 |
| Reasoning 2 | Drop test | 0 |
| Orientation 1 | Drop test | 0 |
| Orientation 2 | Retain as is with 24 items<br>New name ORIENTATION TEST | 10 |
| Orientation 3 | Retain as is with 20 items<br>New name MAP TEST | 12 |
| | Total Test Time | 63 |
| | Est. Administration Time | 36 |
| | Total Time | 99<br>(1 hr. 39 min.) |

subjects, on the average, completed the computer battery in about 94 minutes. For the Concurrent Validation administration, about 65 minutes would be available for the computer battery. Thus, about one-third of the total testing time had to be eliminated without reduction in reliability from some or all of these measures.

To reduce the battery, then, we first examined the reliability estimates for each of the measures. Results from internal consistency estimates indicated that the number of items in some tests could be reduced without reduction in reliability. These measures include Short Term Memory, Perceptual Speed and Accuracy, Target Identification, Cannon Shoot, Target Tracking 1, Target Tracking 2, and Target Shoot. One important outcome of reducing the number of items relates to eliminating rest periods inserted in tests. In other words, for some of the longer tests, we included a rest period to ensure that fatigue would not play a part in test performance. With fewer items, the rest periods could be eliminated, thereby saving time.

Plans for the two reaction time measures include retaining the Simple Reaction Time Test as is, and adding items to the Choice Reaction Time Test to increase the reliability of mean scores. (Note that although the test-retest reliability estimates are low for mean scores on the Simple Reaction Time Test, this test serves as a warm-up for the other reaction time measures and therefore did not require additional items.)

Plans for modifying the Number Memory test differed slightly

from the others. First, this test was viewed as "optional" for inclusion in the final computer battery. That is, this test would be included if time permitted. Second, the test was modified rather extensively from the form in which it was administered at Fort Knox. As noted above, the test contained arithmetic problems that varied in length, with 4, 6, or 8 parts. Analysis of item length indicated that this parameter had very little impact on mean time to perform arithmetic operations. Therefore, we decided to include shorter items consisting of 2, 3, or 4 parts. A final decision about whether or not to include this test in the battery was based on mini-tryouts conducted at the Minneapolis MEPS (N = about 30). These revealed that most subjects completed the total battery in about an hour.

Thus, the final computer test battery included all original tests with most reduced somewhat in length from their Fort Knox form. Table 9 contains a list of the computer tests along with a brief description of the modifications for each. As a final note about the computer battery, it is important to report that several cosmetic changes were made. For example, we again revised test instructions to ensure clarity and uniformity. In addition, the program itself was modified extensively to reduce the time required to operate the test files and to process and store test scores. The end result is a series of tests that function more as a battery of tests rather than as a collection of tests.

Table 9

Changes in Computer Test Battery Cognitive and Psychomotor Measures

| Test Name | Changes from Fort Knox to Summer 1985 Battery |
|---|---|
| COGNITIVE/PERCEPTUAL TESTS | |
| Demographics | Eliminate typing, race, & age items. Retain SSN & video experience items. |
| Simple Reaction Time | No changes. |
| Choice Reaction Time | Increase number of items from 15 to 30. |
| Perceptual Speed & Accuracy | Reduce items from 48 to 36. Eliminate word items. |
| Target Identification | Reduce items from 48 to 36. Eliminate moving items. Allow stimuli to appear at more angles of rotation. |
| Short Term Memory | Reduce items from 48 to 36. Establish a single item presentation and probe delay period. |
| Cannon Shoot | Reduce items form 48 to 36. |
| Number Memory | Reduce items from 27 to 18. Shorten item strings. Eliminate item part delay periods. |
| PSYCHOMOTOR TESTS | |
| Target Tracking 1 | Reduce items from 27 to 18. Increase item difficulty. |
| Target Tracking 2 | Reduce items from 27 to 18. Increase item difficulty. |
| Target Shoot | Reduce items from 40 to 30 by eliminating the extremely easy and extremely difficult items. |

## References


Christal, R. (1983) Personal Communication (6 June 1983).

Gibson, J. J. (ed.) (1947). Motion picture testing and research. Army Air Forces Aviation Psychology Research Program Reports, 7, Washington, D.C.: Government Printing Office.

Guilford. J. P. & Lacey, J. I. (Eds.) (1947). Printed classification tests. Army Air Forces Aviation Psychology Research Program Reports, 5, Washington, D.C.: Government Printing Office.

Jensen, A. R. (1982). Reaction time and psychometric g. In M. J. Eysenck (Ed.), A model for intelligence, Springer-Verlag.

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1982). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10: 1981 Army applicant sample. (Technical Report 581). Alexandria, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences.

Keyes, M. A. (1985, in press) A review of the relationship between reaction time and mental ability. Minneapolis, Minnesota: Personnel Decisions Research Institute.

McHenry, J. J. (1985, in press). The validity and potential usefulness of psychomotor ability tests for personnel selection and classifiction. Minneapolis, Minnesota: Personnel Decisions Research Institute.

McHenry, J. J., & McGue, M. K. (1985) Problems, issues, and results in the development of computerized psychomotor measures. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Peterson, N. G. (1985). Overall strategy and methods for expanding the measured predictor space. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Rosse, R. L. (1985). Advantages and problems with using portable computers for personnel management. Paper presented as the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Sternberg, S. (1966). High speed scanning in human memory. Science, 153, 652-654.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. Acta Psychologica. 30, 276-315.

# EXPANDING THE MEASUREMENT OF PREDICTOR SPACE FOR MILITARY ENLISTED JOBS

Hilda Wing, Chair
U. S. Army Research Institute

August 1985

Symposium presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

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

## Expanding the Measurement of Predictor Space for Military Enlisted Jobs

This symposium, "Expanding the measurement of predictor space for military enlisted jobs," presents research accomplished under Project A, Improving the selection, classification, and utilization of Army enlisted personnel. This research is funded by the U.S. Army Research Institute (ARI), Contract No. MDA 903-82-C-0531, to the Human Resources Research Organization (HumRRO), the American Institutes for Research (AIR), and the Personnel Decisions Research Institute (PDRI). Research scientists primarily from PDRI and ARI have participated in the efforts reported here, on developing new and improved predictors to match an expanded set of criterion measures. All statements made here are those of the authors and do not necessarily express the official opinion or policies of the U.S. Army Research Institute or the Department of the Army.

ARI and Project A have been well represented at this convention, with a variety of workshops, papers and symposia. Those of you who visited the ARI booth in the Exhibition Hall may have had the opportunity to see and perhaps play with our computer demo, which provides a brief example of each of our new computerized measures. While the computer tests are perhaps the centerpiece of our current predictor development efforts, they are far from being all of it. We are here to tell you about that "all of it."

Norm Peterson will begin by describing the overall strategy we have followed in predictor development. Rod Rosse will comment on some of the practical problems encountered in using computers for tests and test development, then Jeff McHenry will discuss some of the conceptual and psychometric concerns we had when computerizing psychomotor skills. Jody Toquam will follow with a description of how we "added to the ASVAB" with cognitive and perceptual tests, using paper and pencil and the computer. Leaetta Hough will complete the substantive portion with what we call the noncognitive predictors: Paper and pencil measures of vocational interests, biodata, and temperament.

We are very pleased to have Jay Uhlaner as our discussant. Jay has been one of our two Scientific Advisors for predictor development; Lloyd Humphreys is the other. They have made invaluable contributions of both their time and their wisdom to this project.

# DEVELOPMENT OF AN INDEX OF MAXIMUM VALIDITY INCREMENT FOR NEW PREDICTOR MEASURES

Lauress L. Wise
American Institutes for Research

Karen J. Mitchell
U.S. Army Research Institute*

August 1985

Presented at the Annual Meeting of the
American Psychological Association, Los Angeles, California

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

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.

*Dr. Mitchell is now affiliated with the Association of American Medical Colleges

## INTRODUCTION

The research results reported here are part of a large-scale long-term effort to expand and validate the selection and classification battery currently used by the U. S. Army. The overall plan and initial results from this effort, known to the Army as Project A, have been documented in Eaton, et. al. (1983, 1984). This paper describes the development of a new index used as one gage of the potential usefulness of alternative supplements to the Armed Services Vocational Aptitude Battery (ASVAB), the current Army selection and classification instrument.

The new measures developed by Project A will be adopted for operational use only if significant gains in predictive validity are realized. The usefulness of each new measure must, therefore, be evaluated in terms of the degree to which it leads to increases in our ability to predict important performance criteria. At present, however, extensive criterion data have not yet been collected on soldiers for whom new and existing predictor data are available. Indeed, the complete set of criterion constructs has not even been fully defined. The Index of Maximum Validity Increment (MVI) was developed as a tool for use in screening an initial array of potential predictors, prior to the collection of criterion data. At this stage, we can only ask what is range of validity increases associated with each new predictor measure across all possible criterion measures.

It is well known that increases in predictive validity are limited by (a) the reliability of the new measure and (b) the extent of overlap with the existing battery. The MVI Index combines information on the reliabilities of both the new and the existing tests and on the correlations among these measures into a single index - an index which gives the maximum possible increment in criterion variance accounted for (squared validity coefficient).

Before detailing the derivation of the MVI Index, we turn to a general description of the development and field testing of the new measures for which MVI Index values were calculated. This description is then followed by the derivation of the new index. The third section of this paper discusses estimation of MVI Index values for a target population when field test data are available for only a selected sample. The next section presents the results of applying the new index to the particular measures under consideration in Project A. The paper then concludes with a discussion of the general usefulness of the new index and of directions for further development.

## THE DEVELOPMENT OF NEW PREDICTOR MEASURES

The development of alternative predictor measures has been managed by Dr. Norm Peterson of Personnel Decisions Research Institute and Dr. Hilda Wing of the Army Research Institute (ref to 1985 APA papers). During this development, two supplemental batteries were assembled and field tested. The first, designated the Preliminary Battery (PB), was a collection of off-the-shelf tests administered to approximately 10,000 first tour soldiers in four different Military Occupational Specialties (MOS). This battery was assembled early in the project and administered to relatively large samples of soldiers to provide the basis for longitudinal analyses of predictor constructs and subsequent performance criterion measures. The emphasis in assembling the PB was on assessing the usefulness of different predictor constructs. The particular measures included in the PB were not candidates for the eventual new supplemental battery. The PB included cognitive, biographical, vocational interest, and temprament measures.

The PB included eight perceptual-cognitve measures. Five were taken from the Educational Testing Service French Kit (Ekstrom, French, and Harman, 1976), two were taken from the Employee Aptitude Survey by Ruch and Ruch (1980), and one came from the Flanagan Industrial Tests (Flanagan, 1965). MVI values were analyzed for these cognitive measures and also for a combined-gender biographical scales based on Owen's Biographical Questionaire (Owens and Schoenfeldt, 1979).

The PB also included 18 scales from the Air Force Vocational Interest Career Examination (VOICE; Alley & Matthews, 1982) and a variety of temperament scales from the Differential Personality Questionnaire (Tellegen, 1982), the California Psychological Inventory (Gough, 1975), the Rotter I/E scale (Rotter, 1966), and the Personality Research Form (Jackson, 1967). While results for these interest and temperament scales are not reported here, the corresponding measures of these same constructs in the PTB are included in results reported for that battery.

The PB was administered at the beginning of Advanced Individual Training (AIT) or One-Station Unit Training (OSUT) between Fall 1983 and July 1984. Extensive performance data are being collected from these soldiers this summer as part of a larger, concurrent validation of a more extensive supplemental predictor battery.

The second supplemental predictor battery to be developed was designated as the Pilot Trial Battery (PTB). The PTB covers a wider range of predictor constructs in the cognitive, vocational interest, and spatial/psychomotor domains and consists primarily of new measures specifically constructed for predicting performance in Army MOS. The PTB was designed to be a "rough draft" of the eventual new supplemental battery. Roughly 300

soldiers completed the PTB in Fall 1984 as part of a field test of this battery.

MVI estimates were calculated for the PTB tests to inform These measures assess psychomotor ability using tracking and firing tasks. Basic processing speed is indexed using reaction time, perceptial speed, target indentification, and short term memory exercises. Perceptual speed and general accuracy are measured by the tracking and short term memory tasks. Cannon shoot and number memory exercises also appear in the PTB.

Early in 1985, results from the administrations of the PB and the PTB were analyzed to inform decisions about revisions to the supplemental predictor measures prior to full-scale tryout in a concurrent validation study. The development of the PB and PTB was guided by a model of job performance constructs and by an extensive review of the the job performance prediction literature. With only minimal validity information on the new measures, theory and expert judgments played a larger role in revision decisions than did empirical results. There were, however, two key questions addressed by the empirical data on both the PB and PTB. The first question was how reliably each of the predictor constructs was measured by the new batteries. The second question concerned the extent to which the new measures overlapped and were redundant with the existing predictor battery, the ASVAB.

In making revisions for the concurrent validation, a significant reduction in testing time was required. The Pilot Trial Battery required approximately 6 hours to administer. The successor battery, the Trial Battery (TB), had to fit within a 4 hour time block. To inform reduction decisions, we sought to develop a single index that imcorporated both reliability and redundancy information. We now turn to a description of the derivation of this index.

## DERIVATION OF THE MVI INDEX

Our goal in developing the MVI Index, was to derive an expression for the maximum possible increase in squared validity in terms of available information on the reliability of the new and existing measures and the correlations among these measures. In addressing a very similar problem, Flanagan (1959) described a "Potential Unique Validity Coefficient". This coefficient is defined as:

SQRT( REL(Y) - R**2(Y;X)/REL(X) ),

where: Y is a new measure being added to an existing battery,
X is a vector of scores on the existing battery,
REL indicates the reliability of the new measure Y and of a composite, X, of the existing measures, and

R**2(Y;X) is the squared multiple correlation when Y is predicted from the existing vector X.

In the derivation of this coefficient, the reliability of Y provides an upper bound for the squared predictive validity of Y alone. The second term, R**2(Y;X)/REL(X), subtracts an estimate of the variance in Y that is already accounted for by the existing set of predictors. Flanagan describes this second term saying:

> "To translate the multiple-correlation coefficient into an estimate of the overlapping variance, the coefficient is first corrected for attenuation due to errors of measurement in the composite. The corrected coefficient indicates the correlation between the single test and a perfectly reliable composite of the other . . . tests."

In practice, however, an existing battery is not a perfectly reliable measure of the overlapping constructs. A new measure may lead to increased validity by adding to the reliability with which common predictor dimensions are assessed as well as by measuring unique predictor dimensions.

The following notation is used in presenting the derivation of the Index of Maximum Validity Increment:

X is the vector of n scores from the existing battery. It is useful to separate X into "true" and "error" components, denoted Xc and Xe, that reflect the reliability of the existing battery.

(01) X = L'*Xc + Xe

Without any loss of generality, we can assume Xc to be a cannonical decomposition of X, with COV(Xc) equal to an identity matrix. Xc may also have fewer dimensions than X and Xe or it may be of full rank. The usual assumptions are made about zero

correlations between the elements of Xe and Xc and that COV(Xe) = Ex is a diagonal matrix whose elements are 1 minus the reliabilities of the different measures in X. L is then a matrix of the regression coefficients for predicting X from Xc. It is the factor pattern matrix in the factor analysis model.

Y is the new score whose potential for increasing predictive validity we wish to investigate. We can similarly decompose Y into common, unique, and error components:

(02) $Y = Byx'*Xc + Byu*Yu + Ye,$

where Byx is a vector of regression coefficients for predicting Y from the common dimensions, Xc; Yu is the reliable unique component of Y uncorrelated with Xc and scaled so that VAR(Yu) = 1; and Ye is the error component with variance Se**2.

In assessing the maximum possible increase in validity that could be attributed to Y, we wish to define a criterion variable Z such that the increase in the variance accounted for when Y is added is maximized. It is clear that Z will be a function of the components of Y and will not contain error, since this would only detract from the variance accounted for by Y. Thus, we can write Z as:

(03) $Z = Czc'*Xc + Czu*Yu,$

where Czc is a vector of coefficients indicating the weight given to each of the common dimensions and Bzy is a coefficient indicating the weight given to Yu. We seek to calculate the particular set of coefficients, Czc and Czu, that will maximize the contribution of Y to the prediction of variance in Z, and then to express the maximum increment to the squared validity coefficient in terms of easily estimated characteristics of Y and X.

For any criterion Z, we consider the partitioned matrix of predictors P' = (Y|X). The correlations among these predictors may be similarly partitioned:

$$(04) \quad Rpp = \begin{pmatrix} 1 & | & Ryx' \\ \hline Ryx & | & Rxx \end{pmatrix}$$

The squared multiple correlation for predicting Z from X and Y is given by:

(05) $SMC(Z;X,Y) = Rpz'*Rpp^{-}*Rpz$

where $Rpp^{-}$ denotes the inverse of the matrix of predictor correlations, and Rpz is the vector of first-order correlations between the predictors and the criterion. Morrison (1967) gives

an expression for the inverse of a partitioned square matrix which may be used here. This expression may be simplified in the present case by the following definition:

(06) Let Uyx = 1 - Ryx'*Rxx~*Ryx

where Uyx is the proportion of variance in Y (unique and error) not accounted for by X, Ryx'*Rxx~*Ryx being the squared multiple correlation of Y with X. The inverse of the predictor correlation matrix may be written as:

$$(07)\quad Rpp^{\sim}=\begin{pmatrix} Uyx^{\sim} & -Uyx^{\sim}*Ryx'*Rxx^{\sim} \\ -Rxx^{\sim}*Ryx*Uyx^{\sim} & Rxx^{\sim} + Rxx^{\sim}*Ryx*Uyx^{\sim}*Ryx'*Rxx^{\sim} \end{pmatrix}$$

Partitioning Rpz' into (Rzy|Rzx) and combining equations (5) and (7), we get

(08) SMC(Z:X,Y)=Rzx'Rxx~Rzx + (Rzy*Rzy-2*Rzy*Ryx'*Rxx~*Rzx + Rzx'*Rxx~*Ryx*Ryx'*Rxx~*Rzx) / Uyx

which may be further reduced to:

(09) SMC(Z;X,Y)=Rzx'*Rxx~*Rzx + (Rzy-Rzx'*Rxx~*Ryx)**2/Uyx

Since SMC(Z;X) = Rzx'*Rxx~*Rzx, the increase in squared multiple correlation due to adding Y to the predictor set is given by

(10) V2I(Z;Y:X) = (Rzy - Rzx'*Rxx~*Ryx)**2/Uyx

The final step in the derivation of the MVI index is to determine the coefficients Czc and Czu that maximize the expression in (10) and then determine the maximum increment in squared validity at that point.

The following expressions show the expressions in (10) in terms of the decompositions of X, Y, and Z given in equations (1)-(3):

(11) Ryx = L'*Byc

(12) Rzy = Byc'*Czc + Byu*Czu

(13) Rzx = L'*Czc

Substituting equations (11) through (13) into (10) gives:

(14) V2I(Z;Y:X)=(Byc'*Czc+Byu*Czu-Byc'*L*Rxx~*L'*Czc)**2/Uyx

This expression can be further simplified using the following definition:

(15) Let Ux = I - L*Rxx~*L'

In this expression, L*Rxx~*L' is the covariance OLS estimates of Xc based on the observed X.

At the same time, we can express Czu in terms of Czc and the VAR(Z). We can scale Z so that VAR(Z) is 1 giving:

(16) VAR(Z) = 1 = Czc'*Czc + Czu**2

Combining (15) and (16) with (14) leads to:

(17) V2I=(Byc'*Ux*Czc+Byu*SQRT(1-Czc'*Czc) )**2 / Uyx

Let Dzc be the value of Czc for which V2I is maximized. At this point, the gradient of V2I with respect to Czc will be zero. The gradient of V2I may be expressed as:

(18) dV2I/dCzc=2*{Byc'*Ux*Czc + Byu*SQRT(1-Czc'*Czc)}
*{Ux'*Byc-Byu*Czc/SQRT(1-Czc'*Czc)}/Uyx

When the first major factor is zero, the function in (18) is minimized at zero. The maximum value is obtained when the second major factor in (18) is zero. In this case:

(19) Dzc = Ux*Byc*SQRT(1-Dzc'*Dzc)/Byu

In order to obtain an explicit expression for Dzc, we substitute Dzu back into the equation and then solve for Dzu in terms of Dzc using equation (16). Substituting Dzc into (19) gives:

(20) Dzc = Ux*Byc*{Dzu/Byu}

Substituting this expression for Dzc into (16) leads to:

(21) Dzu**2*(1 + Byc'*Ux'*Ux*Byc/Byu**2) = 1

This leads to the following explicit solution for Dzu:

(22) Dzu = Byu / SQRT( Byu**2 + Byc'*U2x*Byc )

Substituting this expression back into (20) gives:

(23) Dzc = Ux*By / SQRT( Byu**2 + Byc'*U2x*Byc )

Note that if the constructs Xc were measured with perfect reliability by X, then the covariance of the predicted values, L*Rxx~*L' would equal the actual covariance of Xc, I, so that Ux would be zero. In this case the incremental validity would be

maximized when Czc=0 and Czu=1. Here only the unique portion of Y would contribute to increases in validity. In general, however, the common portion of Y may also contribute to increases in validity coefficients by adding to the reliability with which the underlying constructs, Xc, are measured.

Substituting the expressions for Dzc and Dzu into (17) gives a value for the maximum increase in squared validity of:

(24) MVI={(Byc'*U2x*Byc+Byu**2) / SQRT(Byc'*U2x*Byc+Byu**2)}**2 / Uyx

where U2x = Ux*Ux. Further simplification leads to:

(25) MVI = (Byc'*U2x*Byc + Byu**2) / Uyx

What remains is to translate the terms in (25) into more commonly estimated characteristics of X and Y. Expanding U2x from its definition and (15) gives:

(26) U2x = I - 2*L*Rxx~*L' + L*Rxx~*L'*L*Rxx~*L'

The term L'*L in the middle of the last expression may be expressed in terms of the correlations among the observed measures X and the error variances of X. From equation (1) and the definitions of Xc and Xe, the correlations among the observed measures are given by:

(27) Rxx = L'*L + Ex

Solving for L'*L and substituting into (26) gives:

(28) U2x = I - 2*L*Rxx~*L' + L*Rxx~*(Rxx-Ex)*Rxx~*L'

which simplifies to:

(29) U2x = I - L*Rxx~*L' - L*Rxx~*Ex*Rxx~*L'

Substituting this expression for U2x into (25) gives:

(30) MVI = (Byc'*Byc - Byc'*L*Rxx*L'*Byc - Byc'*L*Rxx~*Ex*Rxx~*L'*Byc+Byu**2) / Uyx

The following identities translate terms in (30) into more commonly estimated statistics:

(31) REL(Y) = Byc'*Byc + Byu**2

(32) Byx = Ryx'*Rxx~ = Byc'*L*Rxx~

where Byx is the vector of standardized regression coefficients for predicting Y from X, and

(33) $SMC(Y;X) = Byx'*Rxx^{-}*Byx = Byc'*L*Rxx^{-}*L'*Byc$

Substituting these expressions for components of (30) leaves us:

(34) $MVI=(REL(Y)-SMC(Y;X)-Byx'*Ex*Byx) / (1-SMC(Y;X))$

Equation (34) completes our derivation of the Index of Maximum Validity Increment. This statistic is defined in terms of the reliability of Y, the squared multiple correlation from regressing Y on X, and the sum of the error variances of the measures in X weighted by the regression coefficients for predicting Y from X.

## ESTIMATION OF POPULATION MVI INDEX VALUES

In deriving the MVI index, we did not introduce the issue of estimation of population values from sample statistics. The issue of estimation is complicated in the present case by the fact that data are available only on pre-selected samples. Ideally, we would like to know the potential contribution of our new measures for a population of applicants or potential applicants, prior to selection.

The MVI index is the ratio of different statistics which are, themselves, nonlinear functions of observed values. The theoretical derivation of the sampling distribution of this statistic is well beyond the scope of the present effort, particularly since we are concerned with something other than simple random samples from the population. Our approach is to estimate corrected population values for each of the components of the MVI index. We then divide our overall sample into replication subsamples and observe the variation in our population estimates across these replications.

Because of the need for replication samples, we have chosen to use only the larger PB sample (n=8943) in our investigation of estimation accuracy. Within the PB, we have selected two types of measures for this part of the study. The first set of measures are the cognitive tests. These tests are expected to have a high degree of reliability, but also a significant degree of overlap with the existing ASVAB. The second set of measures comes from the Biographical Questionnaire. The composite indices generated from the biographical questions are expected to have much lower levels of overlap with the ASVAB and also a greater degree of variation in their reliability.

In carrying out estimation of MVI values, we have used the 1980 norm sample for the ASVAB as the target population for which MVI values are sought. (See Bock & Moore, 1984 for a description of the 1980 norm sample.) This group, a nationally representative sample of 18-23 year olds, is considered a more stable target than populations of actual military applicants, as the characteristics of applicant groups vary over time as a function of the economic and political climate. Bock and Moore (1984) give ASVAB subtest reliability estimates for the 1980 norm sample. Maier and Truss (1983) give descriptive statistics, including ASVAB subtest correlations, for this same sample, and Mitchell and Hanser (1984) give means and standard deviations for the restandardized subtests for this sample.

In estimating the squared multiple correlations between each of the new tests and the existing battery, we have used the classical multivariate adjustments due to Lawley (1943). As described by Lord and Novick (1968, p 146-148), these adjustments involve (1) substituting known population covariance values for

the selection tests, (2) correcting the correlations between selection tests and other variables using:

(35) $Cyx = Syx*Sxx^{-}*Cxx,$

where Syx and Sxx are covariances from the selected sample, and Cxx are the known covariances for the unselected population, and Cyx are the estimates of the population covariances, and (3) adjusting the covariances among the new variables using:

(36) $Cyy = Syy - Syx*(Sxx^{-} - Sxx^{-}*Cxx*Sxx^{-})*Syx'$

where Syy, Syx, and Sxx, again, represent covariances on the selected sample and Cxx and Cyy represent estimates for the unselected population. The squared multiple correlation for each new test is then estimated by:

(37) $SMC(Y) = Byx*Cxx*Byx' / Cyy,$

where $Byx = Cyx*Cxx^{-} = Syx*Sxx^{-}$.

Estimating population reliabilities for the cognitive tests proved to be problematic. These tests were significantly speeded and data from neither parallel forms nor separately timed halves were available for the PB examinees. We decided to use published reliability estimates for these tests. As expected, these reliabilities tended to be lower than coefficient alpha values estimated from the sample at hand, since the coefficient alpha values included spurious components of correlation due to the speeded nature of the tests. While the published reliabilities were not specifically estimated for the present target population, there was no reason to expect any significant differences betwen the publisher's norming populations and our target population.

The biographical questionaire was not at all speeded, so coefficient alpha reliability estimates were judged appropriate. Population reliability estimates were computed by adjusting the sample coefficient alpha values using the correction for heterogeneity differences given in Lord and Novick (1968, p. 130). The specific adjustment was:

(38) $POP\ REL(Y) = 1 - (1 - SAMP\ REL(Y))*SAMP\ VAR(Y) / POP\ VAR(Y).$

The population variances were estimated using the Lawley correction as described above.

In addition to estimating MVI index values for the two sets of PB measures, we sought to estimate confidence bounds for the MVI index estimates. To do this, we divided the whole sample of 8,598 soldiers into 12 independent using the least significant digits of the encrypted SSN. The resulting samples included

between 710 and 720 soldiers each. We computed MVI index values separately for each sample and then examined the variance in the estimates across the 12 replication samples.

We also sought to test the effectiveness of the adjustments for restriction of range by further attenuating observed variances in the sample data. To accomplish this, we redivided the entire PB sample into discrete levels of ability using the AFQT composite of ASVAB subtests thought to index general cognitive ability. Examinees were rank ordered on this index and divided into 12 discrete ability groups. In the ASVAB reference population, the standard deviations of subtest scores has been set at 10. The equivalent standard deviations for the PB sample is shown in Table 1 along with the average of the ASVAB subtest standard deviations across the different replication samples.

TABLE 1
AVERAGE ASVAB SUBTEST STANDARD DEVIATIONS
ACROSS REPLICATE SAMPLES

| | | REPLICATION SAMPLES | |
|---|---|---|---|
| ASVAB SUBTEST | ENTIRE SAMPLE | EQUIV. SAMPLES | STRATIFIED SAMPLES |
| GENERAL SCIENCE | 8.44 | 8.44 | 6.44 |
| ARITHMETIC REASONING | 7.36 | 7.36 | 4.14 |
| WORK KNOWLEDGE | 7.05 | 7.06 | 4.21 |
| PARAGRAPH COMP | 6.79 | 6.80 | 4.75 |
| NUMERIC OPERATIONS | 6.39 | 6.38 | 5.87 |
| CODING SPEED | 7.03 | 7.03 | 6.78 |
| AUTO SHOP | 8.98 | 9.00 | 8.58 |
| MATH KNOWLEDGE | 7.63 | 7.62 | 5.45 |
| MECHANICAL COMP | 8.54 | 8.53 | 7.54 |
| ELECTRONICS INF | 8.14 | 8.14 | 7.49 |

Table 2 shows the resulting mean MVI Index estimates and the estimates of the standard errors of these means based on the variation in index values across replication samples. In these analysis, we also added a second set of "equivalent" replication samples. This new set included 43 subsamples of roughly 200 soldiers each. We sought to examine the performance of the MVI estimators for samples of this size, since this approximated the size of the samples for which PTB field test data were available. (While the entire PTB sample was roughly 300, not all soldiers completed all measures, so 200 is a more realistic approximation to the number of soldiers for whom particular sets of data are available.)

The standard errors shown in Table 2 are estimates of the standard errors of the means across replications. As such they are estimates of the standard errors of the estimate for the entire sample, assuming that the estimator is at least approximately linear (so that the mean of the estimators for the subsamples is approximately equal to the estimator for the whole sample). The values obtained from the equivalent sample replications (with only moderate range restriction) are acceptably low. The median values for the standard errors from the 700 case sample replications are .010 for the cognitive measures and .006 for the biographical scales. The median values are very nearly identical for the estimates based on the 200 case sample replications. These values compare favorably to the theoretical standard error of a (simple random sample) correlation coefficient with a true value of 0. For samples of this size, the correlation coefficient would have a standard error of .011. As with correlation coefficients, there was some tendency for smaller standard errors to be associated with larger MVI values.

The standard errors estimated from the stratified samples (with greater range restriction) were considerably larger. Here the median values were .061 for the cognitive tests and .017 for the biographical scales. The significantly larger standard errors for the cognitive tests were consistent with the significantly greater reduction in heterogeneity for these measures in comparison to the biographical scales.

The results in Table 2, also suggest a degree of mean bias in the MVI values associated with the degree of restriction in range and with sample size. The estimates based on the 700 case equivalent replication samples are consistently about .01 lower than the estimates based on the whole 8500 case sample. The estimates from the 200 case samples were on average .06 lower than the whole sample estimates. The stratified sample estimates showed more major underestimates for those tests were the range restriction was greatest, with a maximum underestimation of .19 for Choosing a Path.

Table 2
MVI Index Means and Standard Errors
For the Total and Alternative Replication Samples

| MEASURE | TOTAL SAMPLE VALUE | 700 CASE REPLICATION SAMPLES MEAN | 700 CASE REPLICATION SAMPLES S.E. | 200 CASE REPLICATION SAMPLES MEAN | 200 CASE REPLICATION SAMPLES S.E. | 700 CASE STRATIFIED SAMPLES MEAN | 700 CASE STRATIFIED SAMPLES S.E. |
|---|---|---|---|---|---|---|---|
| COGNITIVE TESTS | | | | | | | |
| Visualization | .69 | .68 | .010 | .64 | .009 | .54 | .081 |
| Numerical Reasoning | .55 | .54 | .010 | .48 | .015 | .50 | .060 |
| Choosing a Path | .63 | .62 | .009 | .57 | .011 | .44 | .078 |
| Figure Classification | .71 | .70 | .005 | .64 | .008 | .56 | .055 |
| Following Directions | .47 | .46 | .017 | .39 | .022 | .44 | .070 |
| Flanagan Assembly | .31 | .29 | .021 | .22 | .019 | .26 | .049 |
| Hidden Figures | .77 | .76 | .005 | .72 | .006 | .64 | .058 |
| Map Planning | .61 | .59 | .013 | .54 | .014 | .49 | .061 |
| BIOGRAPHICAL SCALES | | | | | | | |
| Academic Attitude | .72 | .70 | .003 | .67 | .007 | .68 | .009 |
| Cultural-Literary | .51 | .49 | .012 | .44 | .012 | .41 | .022 |
| Parental Control | .75 | .74 | .006 | .70 | .007 | .71 | .020 |
| Atheletic Interest | .71 | .69 | .010 | .65 | .010 | .65 | .017 |
| Sibling Harmony | .66 | .65 | .005 | .60 | .009 | .62 | .026 |
| Independence | .50 | .48 | .013 | .42 | .011 | .45 | .015 |
| Scientific Interest | .80 | .79 | .006 | .75 | .006 | .76 | .012 |
| Parental Closeness | .88 | .86 | .003 | .83 | .004 | .83 | .007 |
| Academic Achievement | .78 | .77 | .002 | .73 | .006 | .73 | .016 |
| Leadership | .81 | .80 | .005 | .76 | .008 | .75 | .013 |
| Sociability | .74 | .73 | .006 | .68 | .007 | .70 | .018 |
| Adjustment | .74 | .73 | .004 | .68 | .007 | .70 | .012 |
| Intellectualism | .44 | .42 | .010 | .37 | .011 | .36 | .022 |
| Shop Classes | .28 | .26 | .011 | .19 | .012 | .20 | .031 |
| Offices Classes | .40 | .38 | .009 | .31 | .016 | .31 | .020 |

## MVI VALUES FOR THE PILOT TRIAL BATTERY

Reliability and redundancy values were calculated for the PTB tests in the manner described above. Estimated reliabilities are alpha coefficients corrected for attenuation using algorithm 38). Table 3 gives estimated MVI values for tests in the Pilot Trial Battery, again, treating the ASVAB subtests as the existing battery. Reliabilities, squared multiple correlations, and the composite error measure for the existing measures (ERRX) are also reported.

TABLE 3
MVI AND RELATED STATISTICS
FOR PILOT TRIAL BATTERY MEASURES

| | POPULATION ESTIMATES | | | | SAMPLE VALUES | |
|---|---|---|---|---|---|---|
| MEASURE | MVI | REL | SMC | ERRX | REL | SMC |
| PSYCHOMOTER AND PERCEPTUAL MEASURES | | | | | | |
| Simple Reaction Time | 0.81 | 0.92 | 0.32 | 0.052 | 0.90 | 0.13 |
| Complex Reaction Time | 0.81 | 0.91 | 0.32 | 0.044 | 0.89 | 0.15 |
| Perc. Speed % Corr. | 0.67 | 0.87 | 0.41 | 0.069 | 0.83 | 0.19 |
| Perc. Speed, Mean Time | 0.87 | 0.96 | 0.14 | 0.075 | 0.96 | 0.11 |
| Perc. Speed, Slope | 0.75 | 0.88 | 0.17 | 0.093 | 0.88 | 0.14 |
| Perc. Speed, Intercept | 0.61 | 0.79 | 0.31 | 0.057 | 0.74 | 0.16 |
| TRacking 1 Mean Error | 0.83 | 0.97 | 0.33 | 0.085 | 0.97 | 0.27 |
| Target ID, % Correct | 0.73 | 0.85 | 0.14 | 0.077 | 0.84 | 0.10 |
| Target ID, Mean Time | 0.88 | 0.96 | 0.23 | 0.057 | 0.96 | 0.20 |
| Tracking 2 Mean Error | 0.87 | 0.97 | 0.25 | 0.066 | 0.97 | 0.22 |
| Short Term Mem, % Corr | 0.61 | 0.81 | 0.42 | 0.038 | 0.72 | 0.14 |
| Short Term Mem, Time | 0.85 | 0.94 | 0.17 | 0.066 | 0.94 | 0.11 |
| Short Term Mem, Slope | 0.45 | 0.58 | 0.18 | 0.032 | 0.52 | 0.07 |
| Short Term Mem, Inter | 0.70 | 0.87 | 0.32 | 0.073 | 0.84 | 0.16 |
| Cannon Shoot Time Error | 0.83 | 0.89 | 0.17 | 0.034 | 0.88 | 0.07 |
| Number Memory % Corr | 0.24 | 0.69 | 0.52 | 0.055 | 0.63 | 0.43 |
| Number Memory, Time | 0.76 | 0.96 | 0.48 | 0.083 | 0.95 | 0.37 |
| Target Shoot, Time | 0.86 | 0.92 | 0.16 | 0.036 | 0.91 | 0.11 |
| Target Shoot, Error | 0.75 | 0.86 | 0.18 | 0.069 | 0.86 | 0.16 |
| BIOGRAPHICAL SCALES (ABLE) | | | | | | |
| Emotional Stability | 0.83 | 0.87 | 0.16 | 0.018 | 0.86 | 0.06 |
| Self Esteem | 0.77 | 0.84 | 0.15 | 0.038 | 0.83 | 0.07 |
| Cooperativeness | 0.72 | 0.78 | 0.08 | 0.037 | 0.77 | 0.06 |
| Conscientiousness | 0.72 | 0.82 | 0.11 | 0.063 | 0.81 | 0.08 |
| Nondelinquency | 0.78 | 0.84 | 0.12 | 0.037 | 0.84 | 0.09 |
| Traditional Values | 0.59 | 0.72 | 0.13 | 0.069 | 0.70 | 0.08 |
| Work Orientation | 0.79 | 0.86 | 0.14 | 0.039 | 0.85 | 0.07 |
| Internal Control | 0.70 | 0.81 | 0.19 | 0.053 | 0.79 | 0.11 |
| Energy Level | 0.79 | 0.86 | 0.16 | 0.042 | 0.85 | 0.07 |
| Dominance | 0.83 | 0.88 | 0.16 | 0.015 | 0.86 | 0.05 |
| Physical Condition | 0.85 | 0.87 | 0.04 | 0.020 | 0.87 | 0.03 |
| Social Desirability | 0.64 | 0.71 | 0.15 | 0.021 | 0.68 | 0.05 |
| Self Knowledge | 0.58 | 0.62 | 0.06 | 0.019 | 0.62 | 0.05 |
| Random Response | 0.48 | 0.60 | 0.18 | 0.033 | 0.56 | 0.09 |
| Poor Impressions | 0.56 | 0.64 | 0.14 | 0.023 | 0.61 | 0.06 |

TABLE 3 (Continued)
MVI AND RELATED STATISTICS
FOR PILOT TRIAL BATTERY MEASURES

| MEASURE | POPULATION ESTIMATES MVI | REL | SMC | ERRX | SAMPLE VALUES REL | SMC |
|---|---|---|---|---|---|---|
| PAPER AND PENCIL COGNITVE TESTS | | | | | | |
| Assemble Objects | 0.50 | 0.83 | 0.53 | 0.066 | 0.79 | 0.42 |
| Object Rotation | 0.77 | 0.88 | 0.36 | 0.035 | 0.86 | 0.22 |
| Shapes Test | 0.71 | 0.86 | 0.40 | 0.034 | 0.82 | 0.23 |
| Maze Test | 0.59 | 0.84 | 0.47 | 0.056 | 0.78 | 0.28 |
| Path Test | 0.65 | 0.87 | 0.51 | 0.042 | 0.82 | 0.32 |
| Reasoning 1 | 0.60 | 0.80 | 0.37 | 0.047 | 0.78 | 0.32 |
| Reasoning 2 | 0.43 | 0.64 | 0.31 | 0.033 | 0.63 | 0.29 |
| Orientation 1 | 0.76 | 0.93 | 0.49 | 0.056 | 0.92 | 0.39 |
| Orientation 2 | 0.77 | 0.91 | 0.47 | 0.037 | 0.89 | 0.33 |
| Orientation 3 | 0.58 | 0.91 | 0.66 | 0.049 | 0.88 | 0.56 |
| INTEREST INVENTORY SCALES (AVOICE) | | | | | | |
| Marksmanship | 0.65 | 0.79 | 0.25 | 0.059 | 0.79 | 0.24 |
| Agriculture | 0.59 | 0.68 | 0.12 | 0.041 | 0.68 | 0.13 |
| Mathematics | 0.76 | 0.82 | 0.09 | 0.036 | 0.82 | 0.09 |
| Aesthetics | 0.65 | 0.77 | 0.15 | 0.072 | 0.77 | 0.15 |
| Leadership | 0.76 | 0.82 | 0.13 | 0.033 | 0.81 | 0.06 |
| Electronic Communic | 0.87 | 0.93 | 0.24 | 0.032 | 0.92 | 0.08 |
| Automated Data Proc. | 0.84 | 0.88 | 0.07 | 0.028 | 0.88 | 0.05 |
| Teacher/Counsellor | 0.79 | 0.82 | 0.06 | 0.023 | 0.82 | 0.06 |
| Drafting | 0.75 | 0.85 | 0.15 | 0.068 | 0.85 | 0.14 |
| Audiographics | 0.77 | 0.84 | 0.16 | 0.034 | 0.82 | 0.07 |
| Armor/Cannon | 0.71 | 0.83 | 0.17 | 0.066 | 0.83 | 0.18 |
| Vehicle/Equip Oper | 0.75 | 0.86 | 0.22 | 0.061 | 0.86 | 0.20 |
| Outdoors | 0.65 | 0.80 | 0.25 | 0.062 | 0.79 | 0.23 |
| Infantry | 0.64 | 0.81 | 0.20 | 0.096 | 0.81 | 0.19 |
| Science/Chem Oper. | 0.86 | 0.89 | 0.09 | 0.022 | 0.89 | 0.07 |
| Supply Admin. | 0.90 | 0.92 | 0.08 | 0.013 | 0.92 | 0.03 |
| Office Admin. | 0.90 | 0.94 | 0.10 | 0.032 | 0.94 | 0.09 |
| Law Enforcement | 0.84 | 0.88 | 0.09 | 0.029 | 0.88 | 0.08 |
| Mechanics | 0.75 | 0.95 | 0.41 | 0.099 | 0.95 | 0.36 |
| Electronics | 0.88 | 0.96 | 0.28 | 0.050 | 0.96 | 0.20 |
| Heavy/Combat Constr. | 0.79 | 0.94 | 0.30 | 0.093 | 0.94 | 0.26 |
| Medical Service | 0.92 | 0.95 | 0.07 | 0.027 | 0.95 | 0.07 |
| Food Service | 0.82 | 0.91 | 0.23 | 0.048 | 0.89 | 0.09 |
| Achievement Compos. | 0.19 | 0.55 | 0.41 | 0.029 | 0.36 | 0.16 |
| Safety Composite | 0.46 | 0.59 | 0.19 | 0.021 | 0.53 | 0.09 |
| Comfort Composite | 0.58 | 0.73 | 0.24 | 0.043 | 0.69 | 0.14 |
| Status Composite | 0.49 | 0.70 | 0.34 | 0.031 | 0.63 | 0.20 |
| Altruism Composite | 0.56 | 0.69 | 0.24 | 0.028 | 0.64 | 0.11 |
| Autonomy Composite | -0.08 | 0.17 | 0.17 | 0.069 | 0.11 | 0.11 |

## SUMMARY AND CONCLUSIONS

The Index of Maximum Validity Increment was shown to provide a rational combination of estimates of the reliability and the degree to which the new test overlaps existing measures. It is important to keep in mind that this index is not a prediction of actual validity increments. The MVI index is intended only to provide an ordering of new tests that places tests that are both reliable and not overly redundant ahead of tests that are deficient in one or both respects. The MVI Index was shown to have a level of stability that is not significantly less than the stability of the squared multiple correlations that measure the redundancy component of the index.

The MVI Index did provide an informative summary of information available on the particular measures included in the Pilot Trial Battery. As expected, the paper-and-pencil cognitive tests did exhibit a fairly high degree of overlap with the existing battery ($R^{**}2 > .3$). As a result, the MVI Index values generally fell at or below .7. The biographical and interest measures, on the other hand, tended to show very little overlap with the existing battery. In the case of these variables, the MVI values closely paralleled the estimates of reliability for the individual composites (generally > .7). The psychomotor and perceptual measures, administered by computer, covered the range between these two extremes. The "percent correct" scores from the computer battery tended to have relatively high dedgrees of overlap with the existing battery and correspondingly low MVI values. The response time measures, however, tended to be both relatively independent of existing ASVAB measures and also highly reliable.

There appear to be at least two avenues for further research on the problem of indexing the potential contribution of a new measure to an existing battery. The first concerns appropriate corrections for shrinkage of the estimates in small samples. The underestimation of population values that seems to be related to both small sample size and more extreme restrictions in range, is primarily related to an overestimation of the overlap component of the index. The introduction of ridge regression coefficients or other standard corrections for spuriousness in squared multiple correlations, appears to be indicated.

A second avenue for further research will be to compare the MVI values for these tests with actual increments in predictive validity for a variety of different criterion measures. Such analyses are scheduled in the processing of the Concurrent Validity data now being collected.

## REFERENCES


Alley, W. E., & Matthews, M. D. (1982). The Vocational Interest Career Examination. Journal of Psychology, 112, 169-193.

Bock, R. D. & Moore, E. G. (1984). Profile of American Youth: Demographic Influences on ASVAB Test Performance. Washington DC: Office of Assistant Secretary of Defense (Manpower, Installations and Logistics).

Eaton, N. K. and Goer, M. H. (1983). Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. Research Note 83-37. Alexandria, Va: U. S. Army Research Institute. Eaton, N. K. (1984). Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1984 Fiscal Year. Research Report 1393. Alexandria VA: U. S Army Research Institute.

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. (1959). Flanagan Aptitude Classification Tests: Technical Report. Chicago: Science Research Associates.

Flanagan: J. C. (1965) Flanagan Industrial Test Manual. Chicago: Science Research Associates.

Gough, H. G. (1975). Manual for the California Psychological Inventory. Paul Alto, CA: Consulting Psychologists Press.

Jackson, D. N. (1967). Personality Research Form manual. Goshen, NY: Research Psychologists Press.

Lawley, (1943)

Lord,F. M. and Novick, M. R. (1968). Statistical Theories of Mental Scores. Reading, Mass: Addison-Wesley.

Maier, M. H. & Truss A. R. (1983). Validity of ASVAB Forms 8, 9, and 10 for Marine Corps Training Couirses: Subtests and Current Composites. Alexandria, VA: Center for Naval Analysis.

Mitchell, K. J. & Hanser, L. M. (1984). The 1980 Youth Population Norms: Enlistment and Occupational Classification Standards in the Army. (Technical Report lxxx). Alexandria, VA: U. S. Army Research Institute.

Morrison, D. F. (1967). Multivariate Statistical Methods. New York: McGraw-Hill.

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. Journal of Applied Psychology Monographs, 64, 569-607

Petersen/Wing (1985 APA papers)

Rotter, J. B. (1966). Generalized expectancier for internal versus external control of reinforcement. Psychological Monographs, 80.

Ruch, F. L., & Ruch, W. W. (1980). Employee Aptitude Survey: Technical Report. Los Angeles: Psychological Services.

Tellegen, A. (1982). Brief Manual for the Differential Personality Questionnaire, Unpublished manuscript, University of Minnesota.

U. S. Army Research Institute (1983). Improving the Selection, Classification and Utilizat8ion of Army Enlisted Personnel. Project A: Research Plan. (Research Report 1332). Alexandria, VA: U. S. Army Research Institute.

# PERSONAL CONSTRUCTS, PERFORMANCE SCHEMATA, AND "FOLK THEORIES" OF SUBORDINATE EFFECTIVENESS: EXPLORATIONS IN AN ARMY OFFICER SAMPLE

Walter C. Borman
Personnel Decisions Research Institute

Manuscript submitted for publication

September 1985

The research was performed under U.S. Army Research Institute Contract MDA903-82-C-0531. This research program (Project A) is a long-term, large-scale effort concerned with improving the selection and classification of enlisted soldiers in the U.S. Army. Views expressed here do not necessarily reflect those of the Army or any other agency of the U.S. Government.

Abstract

This research employs personal construct theory (Kelly, 1955) to explore the content of categories or schemata that might be used in making work performance judgments. Twenty-five experienced U.S. Army officers, focusing on the job of non-commissioned officer (first-line supervisor), generated independently a total of 189 personal work constructs they believe differentiate between effective and ineffective NCOs. The officer subjects numerically defined each of their own 6-10 constructs by rating the similarity between each of these constructs and each of 49 reference performance, ability, and personal characteristics concepts. Correlations were computed between the subject-provided similarity ratings for the constructs, and the 189 x 189 matrix was factor analyzed. Six interpretable content factors were identified (e.g., Technical Proficiency, Organization), with 123 of the 189 constructs from 23 of the 25 subjects loading substantially on these factors. Findings here suggest that a core set of constructs is widely employed by these offices as personal work constructs, but that different officers emphasize different combinations of this core set. The core criterion concepts should be useful in the effort to develop NCO performance categories for second tour Army-wide performance rating scales.

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

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

## Personal Construct Theory

As part of his ambitious psychological theory, Kelly (1955) observed that each person characteristically evolves, for his or her convenience in anticipating events (or other persons' activities), construction systems reflecting his/her personal way of viewing and interpreting these events. That is, individuals develop personal construct systems which they use to judge events and to make predictions about future events. Most important for the present purpose is that some of these categories are imposed on their person perceptions. These interpersonal filters may influence observations and judgments about other people by providing frames-of-reference or sets that make perceivers look for certain kinds of interpersonal

information and interpret this information according to their own constructs (Duck, 1982).

Research and practice utilizing personal construct theory consistently employs as an instrument the Kelly Repertory Grid (Rep Grid) procedures. Kelly's method requires subjects first to identify persons they know who fit certain roles (e.g., mother, best friend, etc.) and then to examine triads of these role persons (e.g., role person 1 and 3 vs. 7, 1 and 7 vs. 3, etc.), describing in their own words how the two persons differ from the third. This is done for as many triads as is desired for the particular application.

Once the personal constructs have been elicited, individuals' category systems can be studied in their own right. For example, in clinical settings, where the theory is applied most often, therapists may use the constructs elicited from a patient to help understand that patient's view of other persons, the kinds of differentiating constructs he or she uses in perceiving his/her interpersonal world (Epting, 1984). In addition, individual patients are sometimes asked to rate their role persons on each of their own personal constructs, and these ratings are then correlated or even factor analyzed (for a particular patient) to assess the structure of the patient's personal construct system (e.g., Widom, 1976). Various interpretive rules of thumb have been developed to help personal construct-oriented therapists to diagnose problem "thinking sets" from the kinds of constructs generated and the structure of the construct interrelationships (Adams-Webber, 1979).

Relationships Between Personal Constructs and Other Cognitive Structures

In this section attempts are made to assess briefly similarities and differences between the various cognitive structures most often attended to in the social cognition literature (e.g., Cantor & Mischel, 1977; Hastie, 1981; Landman & Manis, 1983; Rosch, 1978, Wyer & Srull, 1980) and to evaluate how personal construct systems might be related to these concepts. Schema is first of all a generic term that subsumes several other hypothesized cognitive structure terms. Schemata are thought to be categories and/or knowledge structures that persons use to organize and simplify the complex and varied interpersonal information typically present in a social context. In the social information processing sequence of attention, encoding, retrieval, and evaluation (e.g., Taylor & Crocker, 1981), schemata are used to select and pare down the information being processed. They may even be a biasing feature of interpersonal cognitive activity in that perceivers may process observed behavior according to their schematic category structure, at the expense of processing the behavior actually observed.

Regarding different types of schemata, prototypes are hypothesized structures that highlight modal or typical features of a category (Hastie, 1981). Prototypes can be thought of as good examples of a schema (e.g., George is a perfect example of what I mean by dominant). Stereotypes are categories associated with groups of persons (Hamilton & Gifford, 1976). They tend to have, as well, a more affective component than other kinds of schemata. Impl[illegible]

personality theories (Schneider, Hastorf, & Ellsworth, 1979) are said to describe assumptions individuals make about relationships between traits in people. These theories may or may not accurately reflect how traits actually covary in the population.

How does the concept of personal constructs fit in here? First, the concept is in general very similar to the notion of schemata (Landman & Manis, 1983). Personal construct theory posits that category systems for individuals within a "focus of convenience" (a particular context - for example, a supervisor in a work setting) aid in organizing and simplifying information. Further, the concepts of prototypes and stereotypes are not in any way contradictory to the notion of personal constructs. Personal construct theorists have noted that prototype exemplars for constructs can certainly exist and help in better defining an individual's personal categories (Gara, 1982), and personal construct theory views stereotyping as occurring when a person's construct system in relation to a group lacks "individuation and differentiation" (Adams-Webber, 1979). In effect, everyone in the group is seen as standing about the same on his/her personal constructs. Finally, the aspect of personal constructs that emphasizes structure of the construct system and relationships between a person's different constructs is certainly very similar to the concept of implicit personality theory.

Thus, personal construct theory shares most features of the social cognition literature's schematic processing concepts. As we will see, an advantage to applying personal construct theory to the special case of performance appraisal, in addition to the previous

introduction of schema notions to this area (e.g., Feldman, 1981; Ilgen & Feldman, 1983; Lord, Foti, & Phillips, 1982), is that the Rep Grid arising out of research with personal constructs provides a useful vehicle for eliciting categories that may be useful in helping to understand the performance rating process.

Application of Personal Construct Theory to Performance Rating in Organizations

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

Although individual differences in constructs have been emphasized in past research, there may also be substantial similarity in work-related category systems across, especially, experienced

supervisors. Such similarities may result from many observations of incumbents on the job that lead supervisors to similar views of what constitutes effective and ineffective performance.

The relationships of personal constructs to perceptions of work behavior may be akin to what might be called "folk theories" of work performance (Borman, 1983). Interviews with persons about work on jobs sometimes reveal what appear to be deeply felt and sometimes idiosyncratic "theories" of job performance. Consider these statements: A sales manager says with conviction, "You know what the key to this (sales) job is? Thinking on your feet with customers." And, a first-line supervisor speaks, "Show me a person who comes to work on time and I'll show you a good employee." Concepts such as these can be viewed as elements of folk theories and may reflect raters' category systems that help shape judgments about the effectiveness of individual employees.

Of course, characteristics of the work situation and employees themselves will in part dictate what raters observe and process when viewing work behavior. When a salesperson makes the largest sale in the history of the region, the regional manager rater is highly likely to attend to that piece of performance information no matter what the content of his or her personal constructs might be. Also, other features of the situation that increase the salience of a particular construct will make perceivers' use of that construct more likely (Taylor & Fiske, 1978; Tversky, 1977). An example offered by Feldman (1981) is that race is more likely to be a salient construct when a ratee group has only one black than when it contains all blacks.

In spite of potentially relevant situational and ratee factors, the point to be emphasized here is that there may well be important similarities and differences in raters' personal construct systems related to observing and making judgments about work performance. Specifically, raters who have similar construct systems may tend to focus on like aspects of ratee performance and make similar evaluations of its effectiveness; differences in raters' constructs may lead to variations in the work behavior attended to and subsequently recalled in evaluating performance. Thus, personal construct similarities and differences may provide an inherent source of interrater agreement and disagreement.

Although there has been progress in gaining conceptual understanding of how personal constructs and schemata might impact on person perceptions (e.g., Adams-Webber, 1979; Cantor & Mischel, 1979), interestingly, we know little about what such categories may actually "look like" in, for example, some representative sample of perceivers, target persons, and situations. Thus, in the cognitive processing literature, especially as applied to performance appraisal, little is presented regarding what might constitute the substance or the content of these constructs. One intention of the present study was to use procedures developed in personal construct research to give us a glimpse of the nature of work category schemata. Although the study is directly concerned with personal work constructs and folk theories of work performance, hopefully results will be relevant to the literature on schemata, as well.

## Present Research

Regarding applications of personal construct theory to the rating of job performance, research is needed to (a) determine if raters can report meaningful personal constructs related to effectiveness on jobs, (b) examine individual differences in such constructs, (c) evaluate the stability of these constructs in assessing work behavior in different situations and contexts, and (d) assess the impact of these similarities/differences on observations of work behavior and ratings of work performance.

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

# METHOD

## Subjects

Twenty-five officers in the U.S. Army participated in the research, focusing on the noncommissioned officer (NCO: first-line supervisor) job. All officers had at least two years experience managing NCOs, and some had as many as twenty years experience (M = 8.2). Twenty of the 25 officers had 6-10 years in Army management. The officers were all from different units and had varying specialties

(e.g., combat arms, engineering, intelligence).

Procedures

A variant of the Kelly (1955) Rep Grid was used to elicit personal work constructs from the officers. In this research, officer subjects were asked to think of and record the names of nine NCOs they considered to be effective in their jobs and nine NCOs they considered ineffective in their jobs. Six triad combinations of these 18 role persons were then presented. Three triads consisted of two effective versus one ineffective, and the other three compared two ineffective versus one effective. Each role person appeared in one and only one triad. Subjects were asked (in the two effective vs. one ineffective NCO comparison) to record how the effective NCOs were different from the ineffective NCO; that is, what it was about the effective NCOs that differentiated them from the ineffective NCO. Subjects provided a label and a definition for each of these differentiating constructs. The officers were instructed to record for each triad comparison one most salient distinguishing feature between the effective and ineffective NCOs, even if it turned out to be the same or very similar to a previous construct they had recorded.

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

1. Hardworking--Willing to work as long as necessary to accomplish the job; also concerned about the quality of the job.
2. Trustworthy--Once a job has been assigned there is no need to check on him (her).
3. Courage and Candor--Questions dumb rules and speaks own mind.
4. Priorities--Being able to identify those things that must take precedence over others.
5. Technical Proficiency--Knowledge of job and resources to accomplish mission; knows how to do the job better.
6. Firmness--Ability to control personnel and situations without falling apart.
7. Teacher of Soldiers--Always takes the extra time required to ensure soldiers know their task or mission before moving on.
8. Communicates Well--Communicates well with other soldiers, officers, etc., detailed and to the point, tactful, informative, good grammar.

Figure 1. Example personal work constructs.

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

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

Accordingly, 49 reference dimensions were developed to cover the following domains: (a) personal characteristics and personality traits, (b) cognitive and physical abilities, (c) performance constructs relevant to most or all Army enlisted jobs, and (d) military leadership constructs (see Table 1 for the concept labels).

The personal characteristics/personality traits were identified by reviewing the constructs represented in major personality inventories, as well as taxonomic and factor analytic work done in

personality research (Hough & Kamp, 1984). Sixteen personality attributes appeared to cover this domain. The cognitive and physical abilities emerged from reviews of these constructs (Peterson, 1984; Peterson & Bownas, 1982). The nine cognitive and physical abilities included mechanical and verbal ability and physical coordination. The performance dimensions were identified in a large-scale critical incidents study of enlisted soldier effectiveness (Borman, Motowidlo, & Hanser, 1983). The 12 dimensions reflected a broad effectiveness domain including elements of technical job performance, organizational commitment, and organizational socialization. Finally, 12 leadership dimensions for NCO first-line supervisors were developed in an analysis of the NCO job (Hubein, Kaplan, Miller, Olmstead, & Sharon, 1983). As Table 1 shows, these included administration of personnel, training soldiers, and organizing and controlling resources.

The 49 reference constructs were named and carefully defined. The intention was to have subjects rate on a 5-point scale the similarity between each of their own personal work constructs and each of the reference constructs (where 4 = my construct is very similar to the reference construct and 0 = my construct and the reference construct are completely different in meaning). However, a pilot test of this trait implication procedure indicated that some guidance was needed on the distribution of similarity ratings that officers were to make. Accordingly, based on experience with the pilot test,

a modified forced distribution was developed to serve as a target for subjects. The distribution for individual personal constructs across the 49 reference constructs was: 1-3, 4s (i.e., very similar); 3-5, 3s; 6-10, 2s; 9-13, 1s; and 20-28, 0s.

Officer subjects then used the 5-point scale, along with guidance on the target distribution, to make judgments about the similarity between each of their personal work constructs and each reference construct. Again, following Kelly (1955), the notion here was to obtain the subject's own definition of his or her personal constructs, but in a numerical form that would allow correlational analyses to index similarities and differences in the content of different constructs.

Data Analyses

The focal analysis involved simply correlating the vectors of similarity ratings within and across subjects. To clarify, the number of variables in this analysis was the total number of constructs generated by the 25 subjects (189), and the N of each correlation was the number of reference constructs (49). The 189 x 189 correlation matrix was factor analyzed to explore the patterns of similarities and differences in content of the personal work constructs, both within and across subjects. In this manner, subject and content factors might be identified. For example, a factor with all constructs for an individual officer loading substantially on it would suggest the subject has a highly related set of constructs and a comparatively idiosyncratic work construct system, with his/her constructs unrelated to others' constructs (subject factor). A factor highly interpretable

and having work constructs from several subjects loading on it might, however, indicate a construct held in common across these officers (content factor).

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

## RESULTS

Factor analysis results are summarized in Table 2. The eight-factor solution was selected because of interpretability of factors and a substantial drop in eigenvalues for subsequent factors. Six of the factors are readily interpretable. To provide a richer description of the six content factors, the example constructs in Figure 1 were selected so that the first construct is one that loaded highly (> than .70) on Factor 1, the second loaded highly on Factor 2, etc.

Table 3 shows that Factors 3 and 8 are most like subject factors in that for each of these factors one or two officers have several constructs loading on it and very few of the other officers have any constructs associated with the factors. Each of the six content factors are shared by eight or more officers. Of course, some of the

officers have two to five of their own constructs loading on a single content factor.

Table 3 indicates just how much in common the content factors are across the 25 subjects. Constructs associated with three of the factors are held by the majority of the officers (Initiative/Hard Work, Maturity/Responsibility, and Technical Proficiency), and eleven, eight, and eight officers, respectively, have constructs related to the other three content factors (Supportive Leadership, Assertive Leadership, and Organization).

One way to look at the construct similarities/differences question is to consider the number of constructs loading primarily on the content factors. Table 3 indicates that 123 of the 189 personal work constructs generated (65.1%) have substantial loadings on a content factor, and are thus shared with 7-17 other officer subjects. Of the 66 remaining constructs, 21 (11.1%) loaded on subject factors and 45 (23.8%) had mixed loadings or low communalities.

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

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

2. Idiosyncratic--Loadings for the majority of own constructs are either on an uninterpretable subject factor or show low communalities. (5): Subjects 1, 5, 8, 20, and 24.

3. Narrow Focus--Loadings for at least half the constructs are on content factors, but only one or two factors are represented. (3): Subjects 2, 17, and 25.

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

All but five of the subjects have 50% or more of their constructs loading on content factors. Seventeen officers have three or more content-oriented constructs reflected in their systems (differentiated and differentiated but focused), although nine of these seventeen tend to focus on one or two content areas (the differentiated but focused subjects). Finally, three officers hold constructs in common with other subjects, but they attended to just one or two content areas (the narrow-focus group).

## DISCUSSION

Results of this exploratory study show that managers very knowledgeable about a job can articulate what appear to be substantive categories of subordinate effectiveness on that job. Thus, personal

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

Compared to interpersonal dealings in general, making judgments about people in the performance effectiveness domain may involve fewer possible constructs to consider for successful functioning, and this could lead to greater agreement in construct content. Also, relatively standardized leadership training on the part of military officers might have helped produce the similarities across subjects' construct systems in the present sample.

These observations lead to consideration of the etiology of personal work performance constructs, and also to speculation about what meaning they have for perceiving and interpreting performance information. If analogies to personal construct theory are appropriate, these constructs develop over time based on the manager's personal experiences viewing a variety of incumbents performing their

jobs, and his/her making formal and informal evaluative judgments about this performance. Normal, "healthy" development of a manager's construct system might begin with relatively undifferentiated, experimental categories that may often not serve the manager well. In time, however, the category system is likely to become more differentiated (i.e., more multidimensional), with new or refined categories emerging to enable the manager to make useful distinctions between effective and ineffective subordinates. A "good" construct system in this context yields relevant effectiveness criteria and standards and relatively undistorted pictures of subordinates' performance, thus facilitating wise personnel decisions regarding these individuals. A manager's skill in making such informed decisions raises the probability of his/her survival and success.

As mentioned previously, there has been considerable recent discussion and speculation about the role of schemata in performance appraisal judgments (e.g., Cooper, 1981; Feldman, 1981; Ilgen & Feldman, 1983; Lord, Foti, & Phillips, 1982). Feldman (1981) noted that the choice of schemata or categories to employ in judging others' performance depends on both situational factors (especially various salience factors discussed previously, such as memorable, outstanding performance on the part of ratees) and person or perceiver factors. The present research has focused on these person factors.

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

Future research on personal work construct systems should focus on the stability of these constructs for individual manager raters over time and in different performance situations and on the impact of constructs on perceptions and evaluations of ratee work performance. Regarding the latter, of special interest is the hypothesis that raters who have very different construct systems look for and recall different samples of behavioral information and that they form evaluative judgments about performance based on these different samplings, thus providing an inherent reason for interrater disagreement in ratings. More generally, hopefully this study will open another line of research into the cognitive processes underlying performance judgments.

Finally, the work described here provides possible rating categories that might be used for the second tour Army-wide rating scales to be developed within Project A. Core criterion concepts like Organization and Assertive and Supportive Leadership are apparently employed quite naturally by officers when differentiating between effective and ineffective NCOs. Thus, supervisors of NCOs may feel most comfortable using these concepts as rating categories when asked to generate performance judgments of NCO subordinates. Accordingly, it will be important to consider carefully these core-concepts during the effort to develop second tour Army-wide rating scales.

## References

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

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

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

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

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

Cantor, N., & Mischel, W. (1979). Prototypes in person perception. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 12). New York: Academic Press.

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

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

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

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

Gara, M. A. (1982). Back to basics in personality study - the individual person's own organization of experience: The individuality corollary. In J. C. Mancuso & J. R. Adams-Webber (Eds.), The construing person. New York: Praeger.

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

Hamilton, D. L. & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. Journal of Experimental Social Psychology, 12, 392-407.

Hastie, R. (1981). Schematic principles of human memory. In E. T. Higgins, C. A. Herman, & M. P. Zanna (Eds.), Social cognition: The Ontario Symposium (Vol. 1). Hillsdale, NJ: Erlbaum.

Hough, L. M., & Kamp, J. (1984). Utility of personality assessment: A review and an integration of the literature. Minneapolis, MN: Personnel Decisions Research Institute.

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

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

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

Landman, J., & Manis, M. (1983). Social cognition: Some historical and theoretical perspectives. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 16). pps. 49-123. New York: Academic Press.

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

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

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

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

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

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

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

Schneider, D., Hastorf, A. H., & Ellsworth, P. C. (1979). Person perception (2nd ed.). Reading, MA: Addison-Wesley.

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

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

Taylor, S. E., & Crocker, J. (1981). Schematic bases of social information processing. In E. T. Higgins, C. A. Herman, & M. P. Zanna (Eds.), Social cognition: The Ontario Symposium (Vol. 1). Hillsdale, NJ: Erlbaum.

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

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

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

Wyer, R. S., & Srull, T. K. (1980). Category accessibility: Some theoretical and empirical issues concerning the processing of social stimulus information. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), Social cognition: The Ontario Symposium (Vol. 1). Hillsdale, NJ: Erlbaum.

## Author Notes

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

**Table 1**

Reference Concepts Used for Trait

Implication Similarity Ratings

Personal Characteristics and Personality Traits

1. Energy Level
2. Dominance
3. Self-Confidence
4. Sociability
5. Emotional Stability
6. Cooperativeness
7. Aggression
8. Conscientiousness
9. Persistence
10. Orderliness
11. Originality
12. Reflectiveness
13. Achievement
14. Masculinity
15. Independence
16. Flexibility

Cognitive and Physical Abilities

17. Intelligence
18. Good with Numbers
19. Mechanical Ability

(table continues)

20. Good with Words
21. Physical Coordination
22. Physical Strength
23. Work Orientation
24. Steadiness/Precision
25. Perceptual Speed and Accuracy

Performance Constructs for Enlisted Soldiers

26. Stay out of Trouble
27. Controlling Own Behavior Related to Personal Finances, Drugs/Alcohol, and Aggressive Acts
28. Adhering to Regulations, Orders, and SOP and Displaying Respect for Authority
29. Displaying Honesty and Integrity
30. Maintaining Proper Military Appearance
31. Maintaining Proper Physical Fitness
32. Maintaining Own Equipment
33. Maintaining Living and Work Areas to Army/Unit Standards
34. Exhibiting Technical Knowledge and Skill
35. Showing Initiative and Extra Effort on the Job/Mission/Assignment
36. Attending to Detail on Jobs/Assignments/Equipment Checks
37. Developing Own Job and Soldiering Skills

Military Leadership Constructs

38. Effectively Leading and Providing Instruction to Other Soldiers

(table continues)

39. Supporting Other Unit Members
40. General Unit Administration
41. Administration of Personnel
42. Training Soldiers
43. Supervising
44. Organizing and Controlling Resources
45. Planning
46. Group Development
47. Interpersonal Relations
48. Personal Ethics and Attitudes I
49. Personal Ethics and Attitudes II

**Table 2**

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

| Common Variance Accounted For | Factor | Factor Definition |
|---|---|---|
| 20.7 | 1 | Initiative/Hard Work--Having initiative to tackle jobs; self-starter; working hard and for long hours; dedication to tasks and the job; high energy and action orientation. |
| 12.6 | 2 | Maturity/Responsibility--Being consistently mature, responsible, and dependable; integrity and honesty; "good citizen." |
| 9.2 | 3 | Subject Factor--(Uninterpretable) |
| 7.4 | 4 | Organization--Being well-organized; setting priorities; organizing subordinates and resources. |

(table continues)

| Common Variance Accounted For | Factor | Factor Definition |
|---|---|---|
| 12.3 | 5 | Technical Proficiency--Displaying technical proficiency and competence on job; possessing good job knowledge; knowing where to go for technical information (if needed); learning new concepts quickly and thoroughly. |
| 7.8 | 6 | Assertive Leadership--Working through subordinates to accomplish the mission; being confident and in control of subordinates; inspiring confidence in his/her leadership. |
| 10.5 | 7 | Supportive Leadership--Displaying concern for subordinates; teaching and providing feedback to help subordinates; supporting and guiding soldiers. |
| 7.9 | 8 | Subject Factor--(Uninterpretable) |
| 88.4 | | |

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

Table 3

Summary of Officer Subjects' Personal Construct Systems

| Officer Subj. | Initiative/ Hard Work | Maturity/ Responsi- bility | Uninter- pretable | Organi- zation | Factor[a] Technical Profic- iency | Assertive Leader- ship | Supportive Leader- ship | Uninter- pretable | Mixed Loadings |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | 1 | 6 | | | | | | 1 |
| 2 | | 3 | | | | | 3 | | 3 |
| 3 | 3 | 1 | | | 1 | 1 | | | 1 |
| 4 | 2 | | | | 1 | 1 | 1 | | 1 |
| 5 | | | | | | | | 4 | 2 |
| 6 | 1 | | | 1 | | 3 | | | 1 |
| 7 | 1 | | | 1 | | 1 | | | 3 |
| 8 | | | | | | | | | 7 |
| 9 | 2 | 2 | | | 1 | | 1 | | |
| 10 | 1 | 1 | | | 1 | | 1 | | 2 |
| 11 | 2 | 4 | | 1 | 2 | | | | 1 |
| 12 | 1 | 1 | | 3 | | | 1 | 1 | 2 |
| 13 | 2 | | 1 | | 2 | | 3 | 1 | |
| 14 | 2 | 2 | | 1 | 2 | | 1 | | |
| 15 | 4 | | | 1 | 2 | | | | 1 |
| 16 | 4 | | | | 1 | | 1 | | 4 |
| 17 | 3 | | | 2 | | | | | 3 |
| 18 | | 2 | | | 1 | 1 | | 2 | 1 |
| 19 | 5 | | 1 | | 1 | | 1 | | 1 |
| 20 | | 1 | 4 | | 1 | | | | 2 |
| 21 | 2 | 2 | 1 | | 1 | 1 | | | |
| 22 | 2 | | | | 1 | 1 | 1 | | 1 |
| 23 | 4 | 1 | | 1 | 2 | 1 | | | |
| 24 | | | | | 1 | | 2 | | 5 |
| 25 | 2 | 1 | | | | | | | 3 |
| TOTALS | 43 | 22 | 13 | 11 | 21 | 10 | 16 | 8 | 45 |

[a]The criteria for loading on a factor were, first, that this was the highest loading for the construct and, second, that it was .50 or above.