ARI Research Note 95-42

Developing New Test Selection and Weight Stabilization Techniques for Designing Classification Efficient Composites

Cecil D. Johnson, Joseph Zeidner and Dolores Scholarios

George Washington University

for

Contracting Officer's Representative
Leonard A. White

DTIC
SELECTED
SEP 1 2 1995
G

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

Personnel and Training Systems Research Division
Zita M. Simutis, Director

July 1995

19950911 013

DTIC QUALITY INSPECTED 8

United States Army
Research Institute for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.

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

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

Edgar M. Johnson
Director

Research accomplished under contract
for the Department of the Army

George Washington University

Technical review by

Peter Legree
Dale R. Palmer

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

NOTICES

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

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

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1995, July	FINAL 9/91 - 8/93

4. TITLE AND SUBTITLE
Developing New Test Selection and Weight Stabilization Techniques for Designing Classification Efficient Composites

5. FUNDING NUMBERS
MDA903-91-C-0137
062785A
A791
1001
C02

6. AUTHOR(S)
Cecil D. Johnson, Joseph Zeidner, and Dolores Scholarios

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
George Washington University
Office of Sponsored Research
Suite 601
2121 I Street, NW
Washington, DC 20052

8. PERFORMING ORGANIZATION REPORT NUMBER

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

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARI Research Note 95-42

11. SUPPLEMENTARY NOTES
COR: Leonard A. White

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

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The major goal of this research was to specify a classification-efficient methodology for the construction of assignment composites of optimally selected and weighted tests drawn from a single battery of ASVAB and experimental tests and targeting a job family.

The experiments examine the effects of the number of tests included in a composite, using different figures of merit as the standard for the selection of tests for components and stabilizing test regression weights.

The research approach adopted involves a simulation of the Army selection and classification process using Project A validity data. Comparisons of classification efficiency obtained under each experimental condition are reported in terms of mean predicted performance (MPP).

Findings indicate that five-test composites, tailored to operational job families and selected by a predictive validity index to provide positive weights, can provide an acceptable approximation of the maximum obtainable MPP. The results confirm the predictions that the use of efficient test selection procedures and least square weights for tests in assignment composites can improve the utility of the (Cont.)

14. SUBJECT TERMS		15. NUMBER OF PAGES
Personnel selection and classification	ASVAB	63
Classification and assignment	Army personnel policies (Cont.)	16. PRICE CODE
Differential assignment theory		

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

13. ABSTRACT (Continued)

Army assignment process. The results show that optimal classification provides twice as much gain in predicted performance as gain from selection alone.

14. SUBJECT TERMS (Continued)

Model sampling experiments
Army operational aptitude areas
Army job families
Mean predicted performance
Predictive validity
Project A
SQT
Simulation of personnel selection and classification processes

FOREWORD

This report is one of a series of research efforts designed to improve the selection and classification efficiency of the Armed Services Vocational Aptitude Battery (ASVAB). The research bears directly on theoretical issues critical to the comparison of differential assignment theory (DAT) with alternative theories and approaches. At the same time, the research provides the empirical basis for designing nine classification-efficient test composites for assignment.

DAT principles applicable to the construction of new and improved Aptitude Areas (AA) include: (1) the best test composites for either selection or classification are least squares (LSE) composites; (2) an increase in battery size provides a steady increase in classification efficiency as measured by mean predicted performance (*MPP*); and (3) Brogden's 1959 model of *MPP* provides an approximation of the relationships of the validities (*R*) of LSEs, the intercorrelation (*r*) among these LSEs and *MPP*. The greatest hope for increasing classification efficiency *MPP* from either test selection or from reclustering of jobs into job families is through obtaining a smaller value of *r*. A smaller value of *r* tends to result from an increase in battery size, while the effect of an increase in composite size whose tests are selected from a fixed-size battery is more likely to increase the values of *r*. However, when tests for composites are selected from an experimental pool, the union of the sets of tests in each composite (implied battery size) increases as the number of tests in each composite increases. This relationship is such that the overall effect on *r* is downward and classification efficiency upward as composite size increases.

EDGAR M. JOHNSON
Director

ACKNOWLEDGMENTS

The authors would like to thank the staff of the Selection and Classification Technical Area of ARI for their contributions to this research. The Contracting Officer's Representative for this effort was Dr. Leonard A. White of ARI. The authors also would like to thank Dr. Peter Legree of ARI and Dr. Dale R. Palmer for their comments on the draft report.

DEVELOPING NEW TEST SELECTION AND WEIGHT STABILIZATION TECHNIQUES FOR DESIGNING CLASSIFICATION EFFICIENT COMPOSITES

EXECUTIVE SUMMARY

Requirement:

The major objective of this research is the development and evaluation of a methodology for potentially improving the predicted performance of soldiers optimally assigned to job families, in contrast with the traditional methodology that uses incremental predicted validity as the figure of merit and is based on the goal of improving the prediction of performance across MOS. A two-stage selection and classification model is visualized for the purposes of this study, making the selection stage independent of the latter classification and assignment stage. The research specifies techniques to create tailored job family composites and also provides an empirical basis for the comparison of differential assignment theory with alternative theories.

Procedure:

Experiment 1 examines the effect of: (1) increasing the number of tests included in an assignment composite; (2) using test pools of different sizes and test content as a source of selecting tests; (3) using either an index or differential validity or of predictive validity as a means of selecting tests for composites; and (4) stabilizing test regression weights within composites by the use of positive weights only.

Experiment 2 addresses the operational practicalities of constructing differentially efficient test composites by incorporating different back sample sizes for selecting tests and for calculating regression weights. Additionally, unit test weighting is compared with optimal full least squares weights as a means of stabilizing regression weights.

The research approach adopted involves a simulation of the Army selection and classification process using Project A validity data. Comparisons of classification efficiency obtained under each experimental condition are reported in terms of mean predicted performance (*MPP*).

Findings:

The results confirm the predictions of Johnson, Zeidner, & Scholarios (1990) and Zeidner and Johnson (1991c) that the use of efficient test selection procedures and least squares weights for tests in assignment composites can improve the utility of the Army assignment process. The two experiments reported here show that optimal classification provides twice as much gain in predicted performance as gain from selection alone. Results also show that when predictors are selected separately for job family assignment composites (as contrasted to selecting predictors for a battery of fixed size), an index of predictive validity (PV) is superior to Horst's modified H_d index.

Utilization of Findings:

The issues involved in this study have a number of significant operational implications concerning how samples and sets of predictor variables should be selected for analysis, and how operational job family assignment composites should best be designed. While LSE composites consisting of all nine ASVAB tests provide the maximum obtainable classification efficiency, it appears that five-test composites, tailored to the operational job families and selected to provide only positive weights, can provide an acceptable approximation of the maximum obtainable *MPP*.

In defining operational composites from an existing battery or a new operational battery from an experimental test pool, consideration should be given to combining job families or including additional jobs into a job family when the combined analysis sample size is less than 1,000. Large families should be considered for shredding into two or more homogeneous families to provide higher overall *MPP* and more even quality distribution of personnel across job families.

Differential assignment theory supports the supposition that an improved test selection index reflecting both predictive validity and the average intercorrelation among least squares estimates (LSEs) of predicted performance may eventually be found to be best for classification efficiency.

DEVELOPING NEW TEST SELECTION AND WEIGHT STABILIZATION TECHNIQUES FOR DESIGNING CLASSIFICATION EFFICIENT COMPOSITES

CONTENTS

Page

INTRODUCTION....1

BACKGROUND....3

RESEARCH METHOD....5
- Designated Population....5
- Research Paradigm....7
- Experimental Design....10
- Procedures....14

RESULTS AND FURTHER ANALYSIS....19
- Experiment 1....19
- Experiment 2....32

DISCUSSION....36

THEORETICAL CONCLUSIONS AND OPERATIONAL IMPLICATIONS....40

APPENDIX A....45

REFERENCES....53

LIST OF TABLES

Table 1. ASVAB and experimental Project A predictor measures....6
2. Operational job families & Project A military occupational specialties (MOS)....8
3. Experiment 1: Design A--average mean predicted performance across 30 cross samples using the ASVAB....20
4. Experiment 1: Design A--average mean predicted performance across 30 cross samples using the 29-test batter....21
5. Experiment 1: Design B--average mean predicted performance across 30 cross samples comparing test selection for batteries and composites....25
6. Experiment 1: Design B--average mean predicted performance across 30 cross samples--modifications to $H_d m$....29
7. Experiment 1: Average mean predicted performance across 30 cross samples reflecting mean shrinkage due to sampling error in regression weights and test selection....31
8. Experiment 2: Average mean predicted performance for different analysis sample sizes....33

LIST OF FIGURES

Figure 1. Research paradigm....9
2. Experiment 1 design....11

CONTENTS (Continued)

Page

Figure 3. Experiment 2 design.....13
4. Differences between selection and classification approaches to the design of composites.....37

DEVELOPING NEW TEST SELECTION AND WEIGHT STABILIZATION TECHNIQUES FOR DESIGNING CLASSIFICATION EFFICIENT COMPOSITES

Introduction

The objectives of the present research are both theoretical and practical. First, it provides empirical evidence for the predictions of Differential Assignment Theory (DAT) and bears directly on theoretical issues critical to the comparison of DAT with alternative theories and approaches. Second, the research provides the empirical basis for designing personnel selection and classification systems and relates to acknowledged problems in the U.S. Army operational selection and classification system.

DAT predicts an increase in the mean predicted benefit of systematic selection and classification from the use of tailored test composites for assignment. This prediction has specific implications for the way in which composites are constructed, and also suggests that different approaches may be appropriate for the purposes of selection and classification. Built on the premises of Brogden (1951, 1959) and Horst (1954, 1955), DAT argues for the use of tailored tests in operational test batteries which are selected to maximize differential validity. The theory further predicts a positive relationship between the number of tests in an operational battery and the mean predicted performance (*MPP*) gain when all variables used to assign an applicant group are optimal least squares estimates. Hence, the larger the number of optimally weighted tests in a classification battery the greater should be the gain in performance.

DAT's predictions contrast a mixture of *g* theory and validity generalization concepts, currently the more commonly accepted theory and practice in selection and classification. This theory has endorsed the use of assignment composites comprising tests selected to maximize predictive validity in a back sample and, as a direct consequence, emphasizes a single measure of general cognitive ability (*g*). Theorists who argue that the same measures are appropriate for selection and classification also usually regard the amount of incremental predictive validity over *g* provided by additional measures as the relevant basis for determining if anything other than general cognitive ability is required for the construction of assignment composites. One result has been the acceptance by many investigators of aptitude composites consisting of measures of *g*, and perhaps one or two measures of perceptual or psychomotor ability, as sufficient for classification (e.g., Hunter & Hunter, 1984; Schmidt, Hunter, & Larson, 1988).

Recent model sampling research based on designated population values derived from Project A has shown an increase in the mean criterion score of a single applicant group when test batteries constructed to maximize differential validity, rather than predictive validity, are used as full least squares assignment composites (Johnson, Zeidner, and Scholarios, 1990). This research also showed that larger, more diverse, test batteries resulted in a greater gain than did small batteries, thus contradicting one of the principles on which the use of *g*-based composites for classification is based.

The results of Johnson, et al.'s (1990) research suggest the value of DAT principles for the construction of assignment composites of optimally selected

and weighted tests drawn from a single battery and targeting a job family or smaller group of jobs. Although the more common operational situation calls for the selection of tests from a predetermined operational battery to form job-specific assignment composites, the direct selection of tests from an experimental battery to form test composites could provide an alternative approach. Prior to this study it had not been established whether a test selection index of differential efficiency would provide a gain over an index of predictive validity, or whether the use of additional tests in each composite, as compared to the three-test composites presently used by the Army, would provide higher *MPP*.

The present research expands on the findings of Johnson and Zeidner (1991), Johnson, Zeidner and Leaman (1992), Johnson, Zeidner and Scholarios (1990), Statman (1993), Whetzel (1991) Zeidner and Johnson (1991c), and Zeidner and Johnson (in press) to examine whether an appropriately modified differential validity index provides comparable benefits when used to create tailored assignment test composites specific to each job-family, as compared to the creation of test batteries in which all tests are used to form each assignment composite. It also explores the effect of methods for improving weight stability that have been largely designed to reduce the shrinkage of tailored test composites in cross samples (Hunter, 1986). Some investigators in service laboratories believe they are reducing the instability of test weights across job-family aptitude area composites by use of unit weights. Three alternative methods of weight stabilization are investigated in the two experiments reported here: (1) the use of unit weights, (2) the restriction of regression weights to positive values, and (3) the obtaining of better estimates of the covariances among predictors in the youth population by aggregating covariances across job samples.

Experiment 1 examines the effects of: (1) increasing the number of tests included in a composite (2) using test pools of different size and test content as a source of selecting tests (3) using either predictive validity or differential validity as a standard for the selection of tests for composites and (4) stabilizing test regression weights within composites by the use of positive weights only.

Experiment 2 addresses the operational practicalities of constructing differentially efficient composites by incorporating different back sample sizes for achieving test selection and the calculation of regression weights. Back samples and back validity traditionally refers to using the same sample for performing test selection and/or determining best weights as is used in computing a correlation coefficiency for the same test composites. As in Experiment 1, the effects of using different operational test batteries and of increasing the number of tests in the composite also are examined. In addition, unit test weighting is compared to optimal full least squares weighting as a more direct assessment of the Army's method for stabilizing regression weights, and the full least squares weights themselves are derived both from aggregate and separate (non-aggregated) job families. Each of these methods represents an alternative method of stabilizing regression estimates.

In each study, the DAT research paradigm is used to investigate the effect of the experimental assignment variables on the criterion measure of mean

predicted performance (*MPP*). A typical research paradigm for DAT calls for the generation of an analysis sample based on the parameters of a designated population. The present studies use the empirical data from the Army's Selection and Classification Project (Project A) to define the population. From these population parameters, it is possible to generate synthetic test scores in standard score form that have the same expected covariances and validities as are predicted from empirical samples drawn from the designated population. *MPP* is measured after a personnel selection and classification system has been simulated and all "individuals" selected from the applicant group have been optimally assigned to job families.

Background

Previous model sampling research, drawing on data from the Army's Selection and Classification Project (Project A) to define a designated population, has indicated the potential benefits of capitalizing on differential validity in the creation of assignment composites. Johnson, Zeidner, and Scholarios (1990) provided empirical evidence that tests selected using Horst's (1954) index of differential efficiency (H_d), and further selected from a heterogeneous pool of tests, can exceed the *MPP* provided by the best nine-test battery selected using predictive validity. In a follow-up to this study, Scholarios, Johnson, and Zeidner (in progress) also showed that the H_d battery resulted in higher *MPP* than did the present Armed Services Vocational Aptitude Battery (ASVAB). Furthermore, in both these studies, assignment variables based on a battery of 10 optimally weighted predictors resulted in greater *MPP* than 5-predictor batteries. This finding appears to be quite at odds with prior conclusions reported in the literature that a composite of no more than three or four tests, chosen for their predictive validity, provides the greatest efficiency (measured in terms of incremental validity) for both selection and assignment to multiple jobs (e.g. Hunter, Crosson and Friedman, 1985; Schmidt, Hunter, and Larson, 1988).

Results of this model sampling research are of considerable practical significance given that ASVAB-derived test composites tailored for each job family are utilized in the Army's current selection and classification process. The selection of tests from the ASVAB for use in the current aptitude areas(AAs) has been based primarily on the examination of predictive validity (e.g. Maier and Grafton, 1981). In the existing system, the AAs are essentially used as if selection accomplished in conjunction with a minimum cut score for each MOS were their sole purpose. Evidence suggests that the operational AAs possess reasonable predictive validity for a variety of criterion measures; that is, they are approximately as valid as *g* against the criterion for each job family (McLaughlin, et al., 1984; Hunter, et al., 1985). But, consequently, they are virtually ineffective for differentiating between jobs, i.e., most of the composites are as valid for jobs in other job families as for the ones to which they have been matched (Zeidner, 1987). Indeed, recent model sampling experiments have shown a negative gain in *MPP* from using the AAs currently adopted by the Army in the context of an optimal assignment algorithm (Johnson, Zeidner, and Leaman, 1992; Statman, 1992).

The central theme of the present research bears on the credibility of DAT as an alternative to either *g* theory or specific aptitude theory for designing test composites. One alternative to DAT argues for the sufficiency of a single

measure of *g* to accomplish the prediction of performance for different jobs. The other theory argues for the use of incremental validity for the evaluation of tailored tests in the context of multiple jobs.

In the eyes of many *g* theorists and validity generalization proponents, evidence which indicates other than the sufficiency of *g* is explained away as statistical artifact, such as inflation due to unstable regression weights or other sampling error (Hunter, 1986; Hunter et al., 1985). DAT, by contrast, measures test battery efficiency in terms of *MPP*, and predicts an advantage from using assignment variables (AVs) possessing differential validity (DV) across multiple jobs. This DV can be achieved in several ways; e.g. the use of measures other than *g* which are selected for DV or by optimal weighting of measures to achieve maximum DV results in least squares assignment composites (Johnson & Zeidner, 1991; Zeidner & Johnson, in press).

DAT's predictions for the creation of operational batteries which form least squares assignment composites, using all tests in the battery, have already found empirical support (Johnson, et al., 1990; Scholarios, et al., 1994). Questions remain, however, when one considers the direct selection and weighting of tests for composites for different job families, when that selection is accomplished separately for each job family. The issue of operational importance is how can a new set of more effective test composites be identified as a replacement for the ineffective operational AAs currently used by the Army. The following operationally-relevant questions are addressed by the two experiments reported here.

First, does the gain in *MPP* resulting from the addition of more tests to each test composite drop off as rapidly as most *g*-theorists would predict? Alternatively, is the relationship between the number of tests in a composite and *MPP* approximately the same as the already known relationship between the size of a test battery and the magnitude of *MPP* after optimal assignment to jobs? This research examines how much *MPP* is lost from using three-test, five-test or nine-test composites instead of full least squares (FLS) assignment composites.

Second, does an index of differential validity provide the same gain over an index of predictive validity when selecting tests for specific job family composites rather than for test batteries? Specifically, when tests are being separately selected for each aptitude composite, and the goal is to maximize *MPP*, is a modification of Horst's index of differential validity index (H_d), or alternatively, of predictive validity, the preferred test selection index?

Third, when selecting tests for separate job family composites, it becomes vital to consider the test pool from which tests will be selected and hence the "implied" test battery necessary to achieve selection and classification across all job families using the total set of selected test composites. The operational test battery, ASVAB, consists of nine tests that appear to provide sufficient differential information regarding predicted performance to achieve statistically and practically significant gains in *MPP* after optimal assignment. A better operational battery would be provided by the use of a larger number of tests with a different (i.e. non-cognitive) orientation from the tests contained in the ASVAB. The third question addressed, accordingly, is what is the effect of using either the ASVAB or the more heterogeneous pool of 29 Project A tests

as a source of tests for composites? Size and heterogeneity of test pools are confounded in this experiment.

Fourth, the issue of instability in regression weights can be addressed by exploring alternative methods of test weight stabilization relative to conventional full least squares estimates. Each experiment examines one experimental method for achieving stable weights. Experiment 1 assesses the loss in *MPP* resulting from the requirement that tests in new aptitude composites are all positively weighted and Experiment 2 examines the effects of unit weighting as used in the Army's current composites. Experiment 2 introduces an additional element of weight stabilization. The more common method of computing test weights by making separate use of each job validity sample to compute predictor intercorrelations is compared to the aggregation of predictor covariances across all job validity samples to provide a more stable estimate of the population covariances. In the latter method of estimating regression weights, some consistency between predictor covariance estimates and validity estimates is sacrificed in the hope of achieving increased stabilization of the predictor covariance estimates.

Finally, are the relatively small sizes of empirical back samples used for test selection and the calculation of regression weights adequate to overcome the effects of sampling error in the selection of tests and computation of weights? Experiment 2 is designed to contrast relatively small sample sizes resembling the Project A concurrent validation job samples with larger analysis samples.

Research Method

Designated Population

The predictor and criterion data for the two experiments were derived from the concurrent validation phase of the Army's Selection and Classification Project (Project A). In Project A, samples of soldiers assigned to one of the 19 Military Occupational Specialties (MOS) selected for the study were administered twenty experimental predictors. The nine tests of the ASVAB and criterion measures provided scores for a total of 29 predictor variables and five criterion components (Young, et al., 1990) for all soldiers in these MOS validation samples. The 29 Project A predictors used in the present experiments are shown in Table 1. Only one of the five criterion variables, the "core technical" MOS specific criterion, was used, consistent with Project A research indicating that the job-specific criterion alone benefitted from the use of unique predictor equations for optimal prediction in different jobs (Wise, McHenry, and Campbell, 1990). Appendix A provides a further discussion of criterion issues.

The covariances among the 29 predictor variables and the validities of these 29 predictors against each of the 19 MOS specific criteria were corrected for restriction in range and unreliability of criteria (see Scholarios, 1990 and Johnson, Zeidner, & Leaman, 1992). The correction for restriction in range was accomplished in two stages. First, variances and covariances of the predictor scores were corrected back to the youth population, assuming direct selection effects on the ASVAB tests (for which youth population covariances were available (Mitchell and Hanser, 1984)) and assuming indirect selection effects with respect to the other 20 Project A predictors. A sample covariance matrix among the 29

Table 1
ASVAB and Experimental Project A Predictor Measures

Code	Predictor
	ASVAB tests
GS	General Science
AR	Arithmetic Reasoning
NO	Numerical Operations
CS	Coding Speed
AS	Auto Shop Information
MK	Mathematical Knowledge
MC	Mechanical Comprehension
EI	Electronics Information
VE	Verbal ability
	Project A composite predictors
	Paper-and-pencil spatial composite
SPAT	Spatial composite
	Perceptual-psychomotor composites
CPAC	Complex perceptual accuracy
CPSP	Complex perceptual speed composite
NMSA	Number speed and accuracy
PSYM	Psychomotor composite
SRAC	Simple reaction accuracy composite
SRSP	Simple reaction speed composite
	Job orientation composites
AUTO	Autonomy composite
SUPP	Organizational/Co-worker support
ROUT	Routine composite
	Temperament & biodata composites
ADJU	Adjustment composite
DEPN	Dependability composite
COND	Physical condition composite
SURG	Achievement orientation composite
	Interest composites
AUDI	Audiovisual interest composite
COMB	Combat interest composite
FSER	Food service interest composite
PSER	Protective service interest
TECH	Technical interest composite
MACH	Machinery interest composite

Source: Peterson, et al. (1990)

predictors was obtained by aggregating the covariances from all 19 MOS validity samples. The corrected covariance matrix represented an estimate of the covariance in the youth population. Second, the validities computed in the separate MOS validity samples were corrected using the corrected covariance matrix as the population matrix subject to direct selection. A further correction was applied to these data by Johnson, et al. (1992) because the covariances among the predicted performance scores were not positive semi-definite (see Whetzel (1991) for a full description).

The designated population was defined by the corrected 29 by 29 predictor intercorrelations matrix and the corrected 19 by 29 validity matrix for both restriction in range and criterion reliability. For the purposes of this experiment, one of the 19 MOS was eliminated from the data since its sample size was too small to permit obtaining stable validities. In addition, validities for the nine Army AA composites were calculated by averaging of the validities for the appropriate jobs. Table 2 shows the nine operational job families, with related MOS, used in the present experiments.

Research Paradigm

The DAT research paradigm described in Figure 1 was applied to each experiment. Synthetic scores generated from the parameters of the designated population are used to form the three analysis samples and the thirty cross samples.

The generation of the analysis samples of synthetic scores simulates the drawing of individuals with 29 predictor scores and one criterion score from the designated population to form separate validation samples for each MOS. These generated validity samples are then either used separately (non-aggregated samples), or consolidated into the three aggregated analysis samples from which the predictor intercorrelations and the validity vectors for each MOS are used to accomplish test selection and computation of test weights for assignment variables (AV).

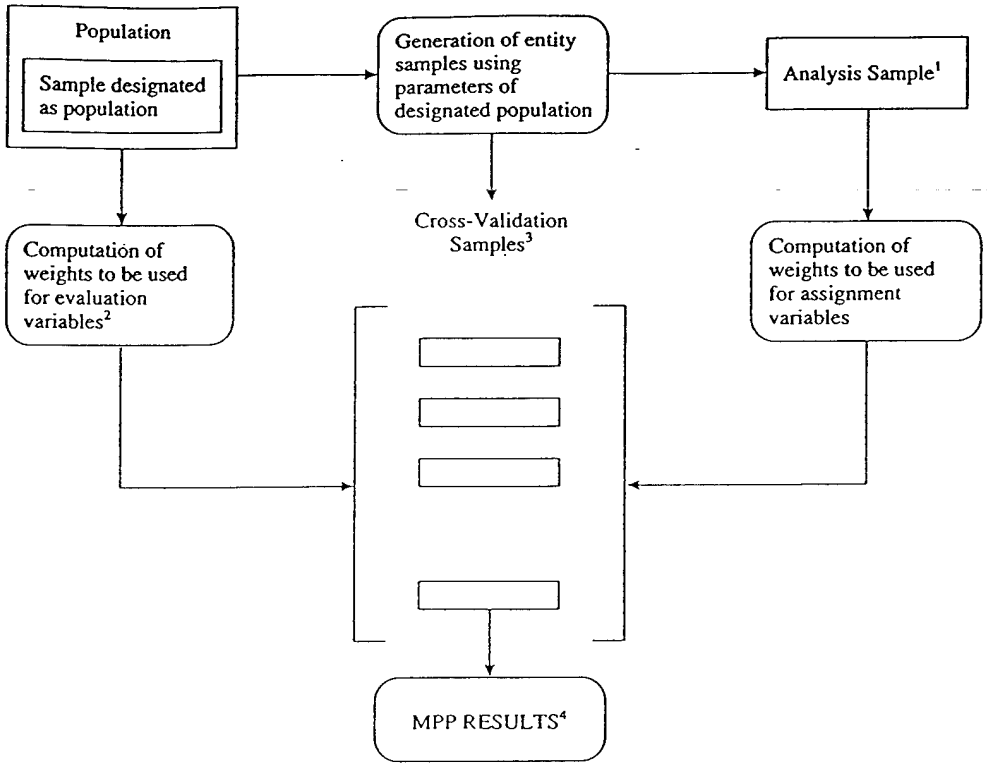
All analysis samples were generated using the parameters of the designated population. Test selection and the computation of regression weights for assignment were accomplished using the predictor intercorrelations and validities of the specified analysis sample.

The model sampling approach assumes that the statistical characteristics of the population of people to be assigned are known. In the present case, the parameters of the corrected Project A empirical sample (the designated population) were assumed to be in standard score form, normally distributed and to have expected correlation coefficients equal to $(\mathbf{R}|\mathbf{V}')'$, where $\mathbf{R}$ is the 29 by 29 matrix of population predictor intercorrelations and $\mathbf{V}$ the 18 by 29 matrix of population predictor validities. Given this information, it was possible to generate samples of artificial test scores with variance/covariance expectations equal to those of the youth population's multivariate normal distribution. These samples of test scores and predicted performance scores were then treated as random samples from the designated population.

Table 2
Operational Job Families & Project A Military Occupational Specialties (MOS)

Operational Job Families	Project A MOS Codes and Names (n=sample size)
Clerical/Administrative (CL)	71L Administrative Specialist (n=427) 76W Petroleum Supply Specialist (n=339) 76Y Unit Supply Specialist (n=444)
Combat (CO)	11B Infantryman (n=491) 12B Combat Engineer (n=544) 19E M49-M60 Armor Crewmember (n=394)
Electronics Repair (EL)	27E TOW/DRAGON Repairer (n=123)
Field Artillery (FA)	13B Cannon Crewmember (n=464)
General Maintenance (GM)	55B Ammunition Specialist (n=203)
Mechanical Maintenance (MM)	63B Light Wheel Vehicle Mechanic (n=478) 67N Utility Helicopter Repairer (n=238)
Operators/Food (OF)	16S Man Portable Air Defense System Crewmember (n=338) 64C Motor Transport Operator (n=507) 94B Food Service Specialist (n=368)
Surveillance/Communication (SC)	31C Single Channel Radio Operator (n=289)
Skilled Technical (ST)	54E Nuclear, Biological & Chemical Specialist (n=340) 91A Medical Specialist (n=392) 95B Military Police (n=597)

Figure 1
Research Paradigm



Notes.

1 Job validation sample sizes equal to those used in Project A first-term concurrent validation study.

2 Evaluation weights computed from Project A empirical sample designated as the population.

3 Sample size of assigned entities number from 200-300; in the aggregate, N numbers in the thousands for each strategy.

4 Predicted performance is computed using the same evaluation variable and same weights for each job across all experimental conditions.)

Optimal assignment to a job family of all "individuals" in each of the 30 cross samples completed the simulation of selection and classification. The entities (artificial individuals), defined as vectors of synthetic test scores, provided 30 samples of entities independent of the analysis sample entities; each such sample is a replication in a repeated measures design. The evaluation (*MPP*) standard scores were computed at the conclusion of each simulation. *MPP* values were produced by first applying full least squares weights from the population parameters to the 29 cross sample test scores possessed by each "individual". The particular vector of population weights applied to each individual's scores depended upon which MOS the "individual" was assigned to during the optimal assignment simulation. The use of population parameters to compute *MPP* scores for the evaluation process maintained a distinction between the assignment variable and evaluation least squares weights, the former of which were computed from the analysis sample. Biases due to correlated error across assignment and evaluation regression parameters were eliminated by the use of the expanded cross validation design used in this study. The more traditional cross validation design would not eliminate this type of bias.

Experimental Design

Experiment 1

Figure 2 summarizes the four facets and corresponding levels of Experiment 1's basic design (Design A) and one additional facet (Design B).

Facet 1 shows the gradual increase in the number of tests in the composite to form three-test, then five-test and finally nine-test best weighted composites. Since three- or four-test aptitude area (AA) composites currently are used by the Army in selection and assignment to jobs, a best weighted three-test baseline appeared to be appropriate. This experiment assessed whether five-test or nine-test composites provided increasing statistically and practically significant gains in *MPP* over a three-test baseline.

Facet 2 permitted examination of two alternative data sources for the selection of tests. The 9 test ASVAB was compared to the 29-test Project A test pool (which includes the 9 ASVAB tests). The comparison was between two different test pools constructed in different ways and with different test content, since the cognitive and non-cognitive experimental predictors of Project A capture different abilities compared to the primarily cognitive ability tests of ASVAB. For the 9 test pool conditions, the experimental design allowed a direct analysis of the effectiveness of alternative approaches of forming composites from the ASVAB as presently used by the Army. Under the 29-test pool conditions, the results provide insight on the potential benefits of diverse test content in enabling selection of composites with increased classification efficiency.

The ASVAB provides a 9-test pool from which the 3 and 5-test composites or batteries can be selected, just as the 29-test pool provides a source for the selection of 3, 5, and 9-test composites or batteries. The implied operational battery is the union of the tests in a set of test composites and represents the set of tests required to be administered to make use of any set of composites. It may well be the size of the implied operational battery, rather than the size

Figure 2
Experiment 1 Design

Design A

(1) Number of tests in composites (3, 5, 9)

(2) Data source from which tests are selected (ASVAB, 29 Project A predictors)

(3) Figure of merit for test selection (modified Horst index of differential efficiency ($H_d m$), predictive validity (PV))

(4) Type of regression weights in composites
   a. least squares using selected tests (i.e. LSE composites)
   b. positive least squares weights obtained by setting negative weights of LSEs to zero; remaining weights are utilized unchanged
   c. positive least squares weights obtained by selecting only those tests whose least squares weights are positive

<u>Design B</u>

(5) Selection of tests for batteries or for composites using facet (4)a only

---

of the composites, that has the greater positive correlation with *MPP*. In this and related studies we refer to the 29-test pool as a battery only when every test in the pool is represented in at least one assignment composite.

Facet 3 provides a contrast between two test selection indices: an index of predictive validity (PV) and a modification of Horst's differential validity index that we call H_dm. Two alternative PV-based indices are appropriate for use as the figures of merit in test selection for a *battery* when the intention is to select personnel for multiple jobs: Horst's (1955) index of absolute prediction (H_a) and Johnson, et al.'s (1990) Max-PSE. These indices lose their distinctiveness, however, when used to select, separately, *composites* for each job family, as in the present experiment. When modified to accomplish the latter objective, the two indices converge to become the index referred to here as PV. With regard to determining the appropriate differential validity index, the point distance index (PDI) has the same relationship to Horst's index (H_d) as does H_a to Max-PSE (see Johnson and Zeidner, 1991). However, a modification of either of these differential validity-type indices to provide an index for use in selecting tests separately for each job family composite produces a more questionable index, one that loses much of the intuitive rationale applicable to the situation where tests are being selected for a single battery. The modification of H_d used here to permit its use in the selection of tests separately for each job family, as noted earlier, is referred to as modified H_d (H_dm).

Facet 4 makes possible the comparison of three methods of test weighting to form test composites. The conventional full least squares method, as employed in previous model sampling experiments, was contrasted to two alternative methods that hopefully might stabilize the regression weights of cross samples by ensuring all weights were positive. The three weighting methods employed were: (a) weights of least squares estimates for sets of all selected tests (least squares estimates); (b) positive least squares weights for the selected tests with negatively weighted tests dropped after computation of the LSEs; and (c) positive least squares weights obtained by imposing a constraint during test selection. This constraint required that the candidate test be rejected if it had a negative weight or caused already selected tests to accrue negative weights.

Facet 5, referred to as Design B, provides a basis of comparison between the results of Johnson, et al.'s (1990) original research and the present study. Although H_d was shown in the original study to be superior to PV for the selection of tests for batteries, its superiority for composites cannot be assumed. In addition, this design enabled a comparison between the performance of the best selected composites and the best selected test battery as alternative methods for creating classification efficient assignment variables. The number of tests required in an operational battery is considerably larger when tests are directly selected for each test composite.

Experiment 2

Figure 3 summarizes the four facets and corresponding levels of Experiment 2. Facet 1 of Experiment 2 corresponds to the size of the analysis sample. Three different analyses samples for carrying out test selection and calculating regression weights for assignment variables were created using three different analysis sample sizes for each job family. A relatively small sample size of

Figure 3
Experiment 2 Design

(1) Analysis sample size for each job family (400, 900, 1600)

(2) Data source from which tests are selected (ASVAB, 29 Project A predictors)

(3) Number of tests in composites (3, 5, 9)

(4) Type of regression weights in composites (3 levels)
   a. least squares using selected tests (i.e. LSE composites) and aggregated job family samples
   b. least squares using selected tests (i.e. LSE composites) and non-aggregated job family samples
   c. unit weights

N=400 provided an approximation of the average size of the Project A empirical MOS samples (see Table 2)--a smaller analysis sample than would be recommended for either test selection or weight computation with respect to defining operational AA test composites.

The data source of tests to be selected makes up Facet 2 of Experiment 2. As in Experiment 1, the ASVAB pool of nine tests was compared to the 29 Project A predictors as an alternative source of tests to be selected.

Facet 3 of Experiment 2 is the size of the test composite. From each pool of tests, three, five or nine tests were selected to form job family composites as in Experiment 1. The nine-test composite formed from the ASVAB was essentially a full least squares composite, and, in the special case of the unit weighting conditions, the nine-test composite using ASVAB tests was a "null" cell and omitted from the experiment. Unlike Experiment 1, only the predictive validity test selection index (PV) was used to select tests for composites.

Facet 4 represents the method of test weighting. In Experiment 2, three different weighting schemes from those of Experiment 1 were analyzed: (a) least squares estimates derived from aggregated job family validity samples of different size (b) least squares estimates derived from non-aggregated job family validity samples of different size and (c) unit weights to represent the current method employed by the Army. The two forms of least squares weighting were used to examine the hypothesis that non-aggregated job samples provide a more accurate and hence stable estimate of regression weights for small samples.

Procedures

Generation of analysis samples

The analysis sample used in Experiment 1 was the same as that used in Johnson, Zeidner, and Leaman (1992). The aggregate analysis sample used to compute predictor intercorrelations and validities was generated to have the same number of predictors (n=29) and "individuals" within each of the 18 job samples (N) as in the concurrent empirical Project A samples (see Table 2). In this approach an m by n matrix of random normal deviates was transformed using a Gramian factor solution (F_u) of the "universe" intercorrelation matrix (R_u) calculated from the corrected Project A empirical data. F_u was used to transform the random normal deviates into test scores for each of the 18 job subsamples. The correlations among the predictors and with the criterion scores for each job provide the intercorrelation and validity matrices, R_{ai} and V_{ai}. The R_{ai} matrices are aggregated across jobs to form R_a and the m separate 1 by n V_{ai} matrices are concatenated to form an m by n matrix, V_a. The analysis sample parameters are provided by the m and n by n super matrix, $[R_a \mid V_a']'$. This analysis sample matrix is the data source for both test selection and the computing of weights for assignment variables.

In Experiment 2, analysis (back) samples of different sizes were formed. Facet 4 of Experiment 2 demanded not only an analysis sample for the aggregated nine job families but also the creation of non-aggregated analysis samples for

N=400 provided an approximation of the average size of the Project A empirical MOS samples (see Table 2)--a smaller analysis sample than would be recommended for either test selection or weight computation with respect to defining operational AA test composites.

The data source of tests to be selected makes up Facet 2 of Experiment 2. As in Experiment 1, the ASVAB pool of nine tests was compared to the 29 Project A predictors as an alternative source of tests to be selected.

Facet 3 of Experiment 2 is the size of the test composite. From each pool of tests, three, five or nine tests were selected to form job family composites as in Experiment 1. The nine-test composite formed from the ASVAB was essentially a full least squares composite, and, in the special case of the unit weighting conditions, the nine-test composite using ASVAB tests was a "null" cell and omitted from the experiment. Unlike Experiment 1, only the predictive validity test selection index (PV) was used to select tests for composites.

Facet 4 represents the method of test weighting. In Experiment 2, three different weighting schemes from those of Experiment 1 were analyzed: (a) least squares estimates derived from aggregated job family validity samples of different size (b) least squares estimates derived from non-aggregated job family validity samples of different size and (c) unit weights to represent the current method employed by the Army. The two forms of least squares weighting were used to examine the hypothesis that non-aggregated job samples provide a more accurate and hence stable estimate of regression weights for small samples.

Procedures

Generation of analysis samples

The analysis sample used in Experiment 1 was the same as that used in Johnson, Zeidner, and Leaman (1992). The aggregate analysis sample used to compute predictor intercorrelations and validities was generated to have the same number of predictors (n=29) and "individuals" within each of the 18 job samples (N) as in the concurrent empirical Project A samples (see Table 2). In this approach an m by n matrix of random normal deviates was transformed using a Gramian factor solution (F_u) of the "universe" intercorrelation matrix (R_u) calculated from the corrected Project A empirical data. F_u was used to transform the random normal deviates into test scores for each of the 18 job subsamples. The correlations among the predictors and with the criterion scores for each job provide the intercorrelation and validity matrices, R_{ai} and V_{ai}. The R_{ai} matrices are aggregated across jobs to form R_a and the m separate 1 by n V_{ai} matrices are concatenated to form an m by n matrix, V_a. The analysis sample parameters are provided by the m and n by n super matrix, $[R_a \mid V_a']'$. This analysis sample matrix is the data source for both test selection and the computing of weights for assignment variables.

In Experiment 2, analysis (back) samples of different sizes were formed. Facet 4 of Experiment 2 demanded not only an analysis sample for the aggregated nine job families but also the creation of non-aggregated analysis samples for

and H_dm required that all rows of the validity matrix, representing all 9 job families, be considered in its computation for each trial test. The PV figure of merit, on the other hand, maximizes the contribution of a trial test to the multiple correlation of previously selected tests and the trial test with respect to the criterion of a specific, separately considered, job criterion. For the separate computation of PV within a given job family, therefore, only the row of the validity matrix representing the appropriate job family was used in the calculation. The formulae for each figure of merit below illustrate this distinction.

$$H_dm_{(m)} = \sum_j^k (a_{ij} - a^*_j)^2$$

where,
- i = the row representing job family m (m=1,...,9);
- j = the column representing the trial test;
- k = the number of previously selected tests plus one trial test;
- a_{ij} = the trial test's orthogonal component in the factor matrix for the i^{th} family;
- a^*_j = $(1/n_m)\Sigma_i^m (a_{ij})$, where n_m equals the number of job families.

$$PV_{(m)} = \sum_j^k (a_{ij})^2$$

where,
- i = the row representing job family m (m=1,...,9);
- j = the column representing the trial test;
- k = the number of previously selected tests plus one trial test;
- a_{ij} = the trial test's orthogonal component in the factor matrix for the i^{th} job family.

A further modification to the test selection procedure was required in Experiment 1 when the tests with positive weights were empirically determined at the test selection phase (facet 4, level 3). Additional constraints and modifications which would avoid the selection of tests with negative LSE regression weights were applied in the selection algorithm.

First, each selected test's semi-partial correlation coefficient, given all previously selected tests, was restricted to a positive value. This coefficient is represented by the a_{ij} coefficients described in the above formulae for H_dm and PV. Thus the j^{th} trial test's orthogonal component for the i^{th} job family criterion (a_{ij}) was considered prior to the computation of the figure of merit for test selection, H_dm or PV. Only trial tests with positive a_{ij} coefficients were retained in the selection process for further consideration.

Second, each selected test was also constrained such that the regression weights applied to each of the already selected tests, as well as the candidate test, were positive, ensuring that the validities of previously selected tests remained higher than each new test and that the intercorrelation between each previously selected test and the new test was sufficiently low. Such a constraint was required to ensure that, as a test was selected, it did not have the effect of turning any previously selected test into a suppressor variable. A suppressor effect would result in some or all previously selected tests having negative regression weights in the eventual combined-variable LSE composite. The implementation of this constraint required a pairwise comparison using the candidate test and each previously selected test in turn. If any of the pairwise comparisons for a candidate test failed to meet this constraint, it was eliminated from consideration. For the t^{th} trial test and the p^{th} previously selected test, the constraint was imposed by:

$$r_{ty} - r_{pt} * r_{py} >= 0.02$$

where, r_{ty} = the validity of the t^{th} trial test for the job family criterion y,
r_{pt} = the correlation coefficient between the t^{th} trial test and the p^{th} previously selected test,
r_{py} = the validity of the p^{th} previously selected test for the job family criterion y.

Finally, for Experiment 1, a non-squared H_dm index was used to rank eligible tests at each test selection phase. For those tests which had satisfied the two constraints, ultimate test selection was based on the largest H_dm before squaring. The use of a non-squared H_dm index was necessary to allow the distinction between negative and positive values in the final comparison of H_dm.

Each of the above constraints was intended to eliminate the occurrence of negative weights for LSEs under the empirically-determined positive weighting conditions of Experiment 1. It was still possible, however, that even after applying the above constraints, negative weights would occur. Furthermore, the goal of selecting 3, 5 or 9 tests could be viewed as the maximum number of tests that could be obtained after constraining eligible tests to positive weights. Under the increased constraints of the modified test selection algorithm, it was likely that some of these targets would not be achieved. It was found, for example, that the H_dm index, while returning the maximum target of 3, 5 or 9 tests from the constrained test selection algorithm, subsequently produced a small number of negative regression weights. In this situation, any negative weights were set to zero in the same way as for the other positive weighting condition of facet three. By contrast, the tests selected by the PV index always returned positive regression weights. However, under some conditions, the desired number of positively weighted tests could not be obtained. In such cases, the test composite was constructed from the reduced, but optimal, set of tests which did satisfy the constraints of the empirical positive weighting condition. In effect, a test which would yield a negative regression weight would instead be given a zero weight.

Generation of cross sample assignment scores

The method for generating the cross samples of test scores and predicted performance scores followed that of the previous model sampling experiments using the corrected Project A data as the designated population (see Johnson, Zeidner, & Scholarios, 1990; Johnson, Zeidner, & Leaman, 1992; Statman, 1992). A simulation of selection and assignment is accomplished within a cross validation design and followed by an evaluation process in which population weights are utilized.

Each of these experiments employed a three-stage procedure for creating samples of assignment variable (AV) scores for use in the simulation of selection and assignment operations. First, random normal deviates were generated to represent the total sample's test scores for each predictor. In the present experiments, thirty cross samples of N=363 were generated, thus providing thirty replications of each experiment for the 363 "individuals". Second, the random normal deviates were transformed into test scores simulating the characteristics of the population from which the samples were assumed to be drawn. The transformation was based on a Gramian factor solution of the population predictor intercorrelations (see Johnson, et al., 1990). Finally, the test scores for these selected "individuals" were used to create AVs derived as LSEs using regression weights calculated from the variable set and analysis sample specific to each of the experimental conditions.

The simulation of the selection process was accomplished using a selection ratio of .70 to eliminate the lower 30% of all "individuals" ranked by their scores on the Armed Forces Qualifications Test (AFQT). In each cross sample, the 363 "individuals" created in the first three steps were reduced to 252.

"Individuals" within the 30 cross samples corresponding to each of the 30 experimental conditions were optimally assigned to job families on the basis of their AVs. A network optimization model was used where the objective function to be maximized was the mean predicted performance (*MPP*) standard score of assigned entities (see Scholarios, 1990 for a more complete description). In each simulation, optimal assignment of 252 entities was accomplished by meeting equal job family quotas of 28 entities in each job family.

In Experiment 1, a single analysis sample provided the appropriate correlation data for calculating AV weights in all conditions. In Experiment 2, the non-aggregated and aggregated job family analysis samples were the source of data for AV weights for each of the three sizes of analysis samples, providing six separate analysis samples.

Computing and Using *MPP* as the Unit of Analysis

The evaluation process calls for the computation of predicted performance scores for each "individual" with respect to the job family to which each person was assigned during the simulation of the assignment process. These predicted performance scores are then averaged across all job families to form an *MPP* score for each replication within each experimental condition. Thirty replications

were accomplished for each condition (cell) in the experimental design (combination of experimental conditions).

Predicted performance scores for each individual and each job family were computed by applying population regression weights to the same test scores as were used in the computation of AVs (using the analysis sample weights). Thus the benefits of a traditional cross validation design are obtained augmented by further safeguards against biasing effects of correlated error that would result from using the same data source to compute weights for AVs and evaluation composites.

The *MPP* value on which experimental conditions were evaluated represented only classification effects. The increase in *MPP* attributable to initial operational selection was computed as a function of selection ratio (SR) and the validity of the selection variable (Naylor and Shine, 1965). As noted above, all simulations in this experiment used a SR of 0.70. The average AFQT validity for the Project A concurrent validation data used here was calculated as .531. Using these parameters yields an expected *MPP* attributable to selection alone of .2623. This constant was subtracted from the *MPP* resulting from all simulations, leaving an *MPP* attributable solely to the classification process. All results presented in this report represent the average *MPP* for 30 replications.

Results and Further Analysis

Experiment 1

Design A

Tables 3 and 4 show the average *MPP* across 30 replications from each experimental condition of Experiment 1. A four-factor repeated measures ANOVA encompassing the full set of replications in both these tables indicate significant main effects for the type of test selection index used ($F_{1,29}$=639.93), the composite size ($F_{1,29}$=864.78), and the weighting method ($F_{1,29}$=277.88) all at $p<.0001$. The facet distinguishing the test pool (the ASVAB or the 29 Project A tests) was not significant at $p<.05$ (F=3.55). However, statistically significant ($p<.0001$) two-way interaction terms for this facet suggested that each experimental assignment variable was behaving differently under different test pools. Further hypothesis testing was conducted separately for the ASVAB and the 29-test battery.

One of the most striking results in both test sets was the performance of Horst's index modified to select job family composites (H_dm) relative to the index of predictive validity (PV). In contrast to Johnson, et al.'s (1990) finding for test selection when the objective was constructing *test batteries*, PV performed consistently better than H_dm for test selection when the objective was *test composites*. Over all 29-test pool conditions, PV provided an average increase of .05 in *MPP* over H_dm ($F_{1,29}$=866.5, $p<.0001$). Overall, selection from the ASVAB showed less differentiation between the two indices, with the PV index giving an average increase of .009 in *MPP* over H_dm over all conditions. H_dm, therefore, appeared to be a less useful test selection method for creating tailored composites directly from an experimental pool, specifically from the 29-test battery.

Table 3

Experiment 1: Design A - Average Mean Predicted Performance across 30 Cross samples using the ASVAB[a]

Test pool	9 test ASVAB					
Test selection index	H_dm			PV		
Number of tests in composite	3	5	9	3	5	9
Test weighting						
Full least squares estimates	.199 (.037)	.230 (.032)	.243 (.035)	.220 (.037)	.234 (.037)	.243 (.035)
LSEs with negative weights set to zero	.181 (.037)	.211 (.036)	.224 (.040)	.209 (.038)	.218 (.041)	.224 (.040)
Positively-weighted LSEs [b]	.208 (.037)	.219 (.035)	.230 (.037)	.214 (.039)	.223 (.038)	.236 (.035)

Notes. [a] Standard deviations for the 30 cross samples are given in parentheses
[b] Positively-weighted LSEs were obtained by selecting tests that would yield positive least squares weights.

Table 4
Experiment 1: Design A - Average Mean Predicted Performance across 30 Cross samples using the 29-test battery

Test pool	29-test pool					
Test selection index	H_dm			PV		
Number of tests in composite	3	5	9	3	5	9
Test weighting						
Full least squares estimates	.173 (.041)	.210 (.038)	.263 (.036)	.221 (.040)	.247 (.042)	.285 (.039)
LSEs with negative weights set to zero	.139 (.046)	.171 (.041)	.212 (.039)	.218 (.041)	.223 (.041)	.242 (.041)
Positively-weighted LSEs [b]	.176 (.039)	.191 (.038)	.204 (.033)	.217 (.041)	.238 (.040)	.266 (.037)

Notes. [a] Standard deviations for the 30 cross samples are given in parentheses
[b] Positively-weighted LSEs were obtained by selecting tests that would yield positive least squares weights.

The contrast between the ASVAB and the expanded 29-test pool also provided some unexpected results. Under the H_dm selection method, the ASVAB *MPP* values were consistently higher (on average, .02 in *MPP*) than those of the 29-test pool ($F_{1,29}$=117.55, p<.0001). This was not the case under the PV method which conformed with the expectation that the larger test pool, which subsumes the smaller pool (battery), would provide greater *MPP* (on average, .02 in *MPP* $F_{1,29}$=41.47, p<.0001).

The remaining comparisons of Design A produced expected results. The 9-test composite resulted in a greater gain in *MPP* than the 5-test composite (.01 for the ASVAB, $F_{1,29}$=195.15, p<.0001; .03 for the 29-test pool, $F_{1,29}$=248.56, p<.0001) which in turn was greater than for the 3-test composites (.02 for the ASVAB, $F_{1,29}$=88.51, p<.0001; .02 for the 29-tests, $F_{1,29}$=501.53, p<.0001). These results for composites are consistent with earlier findings that the larger the test battery the greater is the gain in classification efficiency (CE) (Johnson, et al., 1990). Increases in CE were not asymptotic, over the range evaluated.

The comparison across all weighting conditions supported the expectation that least squares weighting of all tests in the composite is optimal, providing an average increase of .02 in *MPP* over the positive weights obtained at the test selection phase ($F_{1,29}$=138.81, p<.0001) and a .03 increase over the negative weights set to zero ($F_{1,29}$=556.36, p<.0001). Although these differences were statistically significant for both levels of test pools from which tests are selected, they were larger in the 29-test pool conditions. These results show that the least squares weighting provides the best assignment variables, regardless of the number of tests in the composite, the method of test selection, or the test pool from which composites are selected.

A major objective of this study was to determine whether findings relating to the selection of tests for batteries can be duplicated with respect to test composites used as assignment variables. We had hoped to find a modification of H_d that could be used to directly select tests for inclusion in an assignment variable that would provide the same positive effect of using DV for the selection of tests for composites and show the same sensitivity to the increase of the number of tests included in the composites as is found in the selection of tests for batteries.

However, the strong evidence that DV is superior to PV for the selection of tests for inclusion in batteries, from which test composites to be used as assignment variables will be drawn, clearly does not extend to test composites used as assignment variables--when H_dm is utilized as the measure of DV.

The research providing findings in support of the use of DV in selecting tests for batteries all maximize PV in the creation of the test composites used as assignment variables (AVs). Maximizing PVs in AVs in earlier studies provided maximal classification efficiency obtainable from a given battery. Thus, the findings in this study showing the superiority of PV over H_dm for selecting AV tests are not contradictory to the earlier findings with respect to batteries.

Design B

Design A reflects our original intention that Experiment 1 compare the effects of alternative indices (DV vs. PV) used to select tests for assignment composites. Thus, tests were selected from only the 29-test pool. The ASVAB tests by themselves were not also considered as a source. However, it was discovered that when the sample size of the Project A concurrent study was utilized, the *MPP* obtainable from the 9-test pool rivaled that obtained from the 29-test pool. Also, it was discovered that when selecting the test content for job family composites, PV was so superior to H_d that only PV needed further consideration unless we found a convincing way to improve H_dm. The first of these findings caused us to include the 9-test ASVAB pool of tests in all continuing research conducted in this study. The second of these two findings changed our continuing research on H_dm to one of understanding why the DV index was inferior to the PV index under the conditions of design A.

Our discussion of design B results requires us to clarify four important selection-classification system characteristics: (1) test pools from which tests are selected for either batteries or composites; (2) test batteries which under operational conditions must be administered to all recruits; (3) test composites used to classify and assign recruits; and, (4) implied batteries which contain all the tests found in a set of composites. The battery and the implied battery do not differ from an operational point of view. When tests are selected to form a battery in which every composite includes every test in the battery, we refer to this set of tests as simply a battery. However, when tests are directly selected from a pool to form composites, we refer to the total set of selected tests, across all composites, as the "implied battery." The implied battery becomes an operational battery if a set of these directly selected composites becomes operational.

We have not previously felt it necessary to distinguish the impact between size of batteries and size of composites, since we were studying potential classification efficiency where all assignment variables were equal to full least square composites. With this stipulation, our composites and our batteries were the same thing. We, and other colleagues (Statman, 1993; Scholarios et al., 1994; Johnson et al., 1992) have repeatedly shown that the number of predictors contributing to the assignment variables has considerable impact on CE, but we did not try to distinguish between the effect of composite size as compared to battery size. The results of design B permit us to say that increasing battery size (or implied battery size), along with using best weighted composites and increasing the number of job families, constitutes one of the three best ways to increase the efficiency of the Army's classification system.

We began this study with the expectation that we could extend our previous findings regarding the importance of a battery-composite size variable, it now appears that battery or, implied battery, size has more importance in the design of operational systems (and on DAT) than does composite size. While this study was not designed to precisely contrast the contribution to CE of an increase in size of pools, batteries, and composites, it is clear that battery size considerably over shadows these other two system characteristics in its impact on classification efficiency.

The most important distinction in reporting results of design B, as shown in Table 5, is between the indices used to directly select a test battery and the indices used to directly select test batteries from a experimental pool of tests. The latter approach results, indirectly, in a selection of an implied battery. H_d and Max-PSE found in Table 5 and both PDI and H_a also used in previous DAT based research can be used to select a battery but cannot be used to directly select composites. The direct selection of composites using a differential validity concept requires the use of a major modification of H_d, such as H_dm.

Table 5 provides a comparison of two alternative strategies for developing an operational system: (1) selection of batteries followed by the use of all tests in the battery in each AV; and (2) the direct selection of AVs resulting in an operational battery. The first strategy is indicated by the use of indices H_d or Max-PSE, and the second strategy by the use of PV or H_dm.

These comparisons were accomplished using both the ASVAB and 29-test pool for the selection of three-, five- and nine-test composites, using PV and H_dm, and for the selection of batteries containing 3, 5, and 9 tests. The greater size of the operational battery implied by the direct selection of a set of AV composites, as compared to the direct selection of a single battery to be used to make the same assignments, provides a similar advantage in expected *MPP* to the indices used to select composites. Table 5 results do not support the premise that the 29-test pool is demonstrably superior to the 9-test pool (ASVAB) when equal sized operational batteries are being compared. Its contents indicate some consistency with the results of Design A. There was also an increase in *MPP* from the three-test to the five-test composite ($F_{1,29}$=1115.17 and $F_{1,29}$=520.82, p<.0001 for the ASVAB and 29-test pool respectively) and from the five-test composite to the nine-test composite ($F_{1,29}$=127.66 and $F_{1,29}$=879.13, p<.0001 for the ASVAB and 29-test pool respectively).

Most important for interpreting the results of Design A, H_d performs on average .03 in *MPP* better than the predictive validity index (Max-PSE) when tests are selected for batteries in which all test composites are LSEs containing all tests in the battery (i.e., FLS composites). This is the case both for selection from the ASVAB ($F_{1,29}$=133.51, p<.0001) and from the 29-test pool ($F_{1,29}$=55.33, p<.0001) and is consistent with the findings of Johnson, et al. (1990) relating to the selection of tests for batteries. Our DV index, H_dm, does not exhibit this superiority over PV. When tests are directly selected for composites to form implied batteries, increasing battery size for a given composite size, the larger battery, as before, provides the more *MPP*. However, a potential conflict between two proposed DAT principles was considered in this study. These two principles are as follows: (1) tests for composites to be used as AVs should be selected and weighted to maximize PV; and, (2) tests for forming batteries should be selected using a DV index. This conflict, when the DV index is represented by H_dm, is in Experiment 1 resolved in favor of the first of these two DAT principles.

Table 5 findings permit the comparison of benefits provided by selecting batteries vs. composites using a PV index for both strategies. This comparison can be made for same sized composites, but a larger battery is thus implied for the strategy involving direct selection of composites, assuring a larger *MPP* for the index used to select composites. The gain in *MPP* resulting from directly

Table 5

Experiment 1: Design B - Average Mean Predicted Performance across 30 Cross samples comparing Test Selection for Batteries and Composites[a]

Test pool	ASVAB			29-test pool		
Number of tests in composite	3	5	9	3	5	9
Test selection index						
H_dm - composite (Design A)	.199 (.040)	.230 (.032)	.243 (.035)	.173 (.041)	.210 (.038)	.263 (.036)
PV - composite (Design A)	.220 (.037)	.234 (.039)	.243 (.035)	.221 (.040)	.247 (.042)	.285 (.039)
H_d - battery [b]	.188 (.040)	.229 (.036)	.242 (.035)	.170 (.038)	.209 (.041)	.239 (.041)
Max-PSE - battery [c]	.133 (.044)	.206 (.035)	.242 (.035)	.150 (.039)	.173 (.042)	.224 (.040)

Notes. All conditions are full least squares estimates

[a] Standard deviations for the 30 cross samples are given in parentheses

[b] H_d represents Horst's (1954) original index of differential efficiency.

[c] Max-PSE (maximizing Potential Selection Efficiency) maximizes the average validity of the composite.

selecting tests for composites from the ASVAB, using a PV index for both strategies, is .087 for 3-test composites and .028 for 5-test composites. This gain is much less when DV indices are used, as when H_dm is compared to H_d, .011 and .001. Thus, considering only the ASVAB, despite the confounding of the effects of using H_d vs. H_dm with the effects of implied battery size, it is clear that something else than battery size is at work here. H_dm is not an effective index.

The gain in *MPP* resulting from directly selecting tests for composites from the 29-test pool, using a PV index for both strategies, is .071 for 3-test composites, .074 for 5-test composites, and .061 for 9-test composites. Again, this gain is much less when DV indices are used, yielding differences of .003, .001 and .024 for 3-, 5- and 9-test composites.

Further analysis was undertaken to explore the unexpected results of Design B. First, the loss in effectiveness in H_dm relative to H_d could be explained by its lack of theoretical precision as a measure of differential efficiency for creating composites as contrasted to batteries. Horst's (1954) definition of the DV index assumed the presence of the same tests in all composites forming the least squares estimates composite and a predetermined operational battery size. Hence the effectiveness of H_dm as a test selection index for forming composites may depend on increasing the overlap in tests selected for different families while retaining differential test weights across families.

Further examination of the regression weights produced by the H_dm conditions in this experiment revealed that, particularly in the 29-test pool conditions, the test selected first often had near zero weights for the family for which they were selected, by the time all k tests were selected, and thus made little differential contribution. In some cases, there were no differential weights at all to produce the required relative effect on different job families (i.e., there was little overlap of tests selected and the regression weights for earlier selected tests, those yielding the highest H_dm, were close to zero, the implied weight given to non-selected tests; hence weights were not differential across job families).

An extensive literature exists on the merits of building test batteries or composites using a sequential selection of tests with the objective of maximizing selection efficiency in terms of a single criterion variable. The sequential process proceeds one test at a time--either starting from zero (accretion) or from the total experimental test pool (deletion). The literature regarding the selection of tests to maximize a figure of merit based on multiple jobs, each with its own criterion variable is more sparse. Horst published three articles on this specific topic. One article relates to maximizing differential validity by accretion (Horst, 1954) and a second article relates to maximizing absolute validity by accretion (Horst, 1955). A third article relates to maximizing both differential and absolute validity by deletion (Horst and MacEwan, 1960). This literature is discussed in greater detail in Johnson and Zeidner (1991).

In this study the investigators were initially committed to use of the sequential accretion method to select tests for either composites or batteries. In the accretion method the "best" test is selected, then additional "best" tests, one at a time, while always retaining previously selected tests. However,

the deletion method was also used in Experiment 2 in conjunction with a preliminary application of the accretion method. The deletion method commences with a test composite containing all the tests in the pool, or as in this study, with all the tests selected by another method. Each deletion-test selection step results in the removal of one test from the existing pool, with the reduced pool eventually becoming the "best" composite.

The literature, in which the tests selected by accretion and deletion methods from the same test pool are compared, indicates that these two approaches often result in the selection of surprisingly different sets of tests (where test pools are at least moderately large) while providing essentially equal validities (Burket, 1964). Thus, for a given test pool, the union of the "best" 3 tests selected by accretion and the "best" 3 tests selected by deletion provide a set of n tests (n ranging from 3 to 6 depending on the overlap) that would be expected to be better than the "best" n tests selected entirely by either accretion or deletion.

Using an extension of the above logic, a combined use of the accretion and deletion methods should yield higher *MPP* than sets of AVs (assignment variables) selected by only one of the two methods, regardless of whether the figure of merit is a PV or DV index. In this study, the combination of accretion and deletion concepts in the same algorithm was investigated only in conjunction with a DV test selection index (i.e., H_dm).

It appears that the modification made to Horst's DV index to permit the forming of composites was, for a number of reasons, an inadequate index for maximizing classification efficiency resulting from the direct selection from a test pool of a set of equal sized test composites. Various hypotheses were developed to explain why a DV index (H_dm) did not maintain its superiority over the PV index with respect to the *MPP* resulting from a classification process, when separate test selections for each job family were made. These hypotheses provided the rationale for devising further modifications of H_dm in the hope of increasing the classification efficiency resulting from use of this index to select tests for composites. The authors feel that such modifications were worth considering because the earlier finding, as already noted, showed the considerable superiority of H_d over PV for selecting tests for battery rather than for composites (Scholarios, et al., in press).

Hypothesis 1: *An increase in test overlap across job families increases the differential effect of each test in the final set of best weighted test composites and results in greater MPP.* An increase in test overlap was accomplished by a two-stage version of H_dm where a second test selection was

performed against the pool of tests that had been selected in the first stage and had been present in at least two of the nine job families.[1]

The differential validity test selection indices, including H_d, PDI, and H_dm are derived using a sequentially constructed triangular factor analysis solution of the predictor intercorrelation matrix extended to the predicted performance variables. For H_d, the selected candidate test is the one which provides the highest average squared differences of the job family factor coefficient (the one for which selection is being accomplished) from the mean of the 9 loadings of the job family predicted performance variables on the same factor. H_dm differs from H_d in that only one squared difference, instead of the average across the job families, is considered in selecting the "best" test. Before the squaring occurs in the computation of H_dm, this difference can have either a positive or negative sign.

Hypothesis 2: Constraining the H_dm test selection process to the consideration of positive signed differences (as defined above) will provide a set of tests that will provide more potential classification efficiency than will the unconstrained H_dm index. It is believed that the use of this constraint will reduce the number of small, almost zero, regression weights for the first or second tests selected. If this reduction occurs, the LSEs in which the independent variables are selected using the constrained H_dm should provide higher *MPP*s.

Hypothesis 3: A two-step test selection process which selects n + 2 tests by accretion then selects two tests for deletion (to select the best n tests where n equals either 5 or 7) will increase the stability of regression weights resulting in greater MPP.

Table 6 compares the *MPP* scores obtained from assignment of entities using composites selected by the original Design A H_dm (one-stage test selection) with two-stage, constrained and "deletion" versions of H_dm. Indices 1 and 2 represent the unmodified H_dm index from one stage test selection (see also Table 4) and the unmodified H_dm index with two-stage test selection. Indices 3 and 4 apply the constraint, and indices 5 and 6 apply the deletion process to the original indices 1 and 2. For all sizes of composites and for both one-stage and two-stage selection, the constraint resulted in an average decrease of .025 in *MPP*. On application of the deletion process, however, there was an overall average increase of .02 in *MPP* from the original H_dm indices. The greatest *MPP* values were obtained from the two-stage H_dm with deletion (index 6 in Table 6), and, as before, from the 9-test composites.

---

[1] Two-stage test selection was not performed for PV test selection as it would provide no added value to the classification efficiency of resulting composites. When the column variance of the trial tests are evaluated using tests which have already been selected for assigning to other jobs (i.e., for classification), H_dm becomes a more effective figure of merit for selecting tests to be used in a particular job composite. Thus, the presence of tests which are unselected for other job families would detract from the effectiveness of H_dm. With the PV figure of merit, unselected tests have no effect when selection is for one specified job family, since only the tests already selected for a particular job have any effect on the trial test's evaluation.

Table 6

Experiment 1: Design B - Average Mean Predicted Performance across 30 Cross samples - Modifications to H_dm [a] [b]

Test Pool	29-test pool		
Number of Tests in Composite	**3**	**5**	**9**
H_dm Index			
1. H_dm - one stage (Design A)	.173	.210	.263
	(.041)	(.038)	(.036)
2. H_dm - two stage (Design B)	.178	.220	.266
	(.035)	(.032)	(.039)
3. H_dm - one stage/constrained [c]	.132	.201	.255
	(.040)	(.040)	(.037)
4. H_dm - two stage/constrained [c]	.144	.184	.258
	(.044)	(.037)	(.038)
5. H_dm - one stage/with deletion [d]	.208	.232	.280
	(.038)	(.039)	(.033)
6. H_dm - two stage/with deletion [d]	.213	.244	.281
	(.036)	(.036)	(.036)

Notes [a] Standard deviations for the 30 cross samples are given in parentheses

[b] All conditions are least squares estimates

[c] The first 2 tests in the 3-test composites were constrained to have a positive difference between the factor coefficient loading of the job family and the mean of the loadings of all job families on this same factor (see hypothesis 2).

[d] The "deletion" process was initiated from the "best" 5, 6, and 11 tests for the 3-test, 5-test and 9-test composites, respectively. For each condition, the tests with the lowest weight were dropped from the implied battery from which the final composites were created. At each deletion stage, the composite weights were recalculated.

Referring again to Table 4, the greatest classification efficiency in Design A was provided by the LSE assignment composite selected by PV from the 29-test pool (.221 in *MPP* for the 3-test composite, .247 for the 5-test composite, and .285 for the 9-test composite). No other condition evaluated in this experiment produced higher *MPP* values. Only by increasing the test overlap across job families during test selection and by deleting low and unstable weights from the composites (as in index 6, Table 6), was the classification efficiency from $H_d m$ test selection increased to come within .004 in *MPP* of the PV 9-test composite.

A second unexpected result from Design B was that the selection of the best 9-test battery from the 29-test pool was inferior to the *a priori* ASVAB battery. This was the case using both H_d (.239) and Max-PSE (.224) as the test selection indices in the comparison with the ASVAB (.242). Some explanation for this finding was required since the *a priori* ASVAB is subsumed within the 29-test pool and therefore, in a back sample, cannot provide greater *MPP* than the "best" 9-tests selected from the same test pool. A further hypothesis examined whether the apparent superiority of the ASVAB in the present experiment's cross samples was influenced by the sampling error introduced by conducting test selection and computing assignment regression weights on the analysis sample.

Hypothesis 4: Controlling for sampling error in test selection and/or the calculation of regression weights for assignment will increase the MPP provided by the best 9-test battery selected from the 29-test pool sufficiently to provide superiority over the ASVAB.

Further simulations were conducted to permit the comparison of the ASVAB operational battery and the best 9-test battery selected from the 29-test pool using H_d (as in Johnson, et al., 1990). The 29-test pool conditions could be freed of sampling error in either or both the test selection process and/or the computation of assignment weights by using the population, rather than analysis sample, predictor intercorrelations and validities in the analysis process. The ASVAB condition, using an *a priori* battery, is always free of the effects of test selection sampling errors and can be free of sampling error in the regression weights by using the population parameters to compute regression weights. Table 7 shows the average *MPP* standard scores for six conditions. Two of these cells represent the *MPP*s obtained from Design B (Table 5) where no sampling error effects were present (i.e., test selection and computation of regression weights were based on the analysis sample: *MPP* equals .239 for the best 9-test battery and .242 for the ASVAB. The primary result of interest was the effect of shrinkage in *MPP*, measured in the cross samples, attributable to capitalizing on test selection and regression sampling error in the back sample and the effect of this when comparing the *a priori* ASVAB to the best 9-test full least squares composites.

The superiority of the ASVAB over the best H_d-selected 9-test battery shown in Design B disappeared when test selection error in H_d selection alone was absent, resulting in a .03 *MPP* gain over the ASVAB. When regression, but not test selection, error alone was absent, in both conditions the best 9-test battery gave a .01 gain in *MPP* over that provided by the ASVAB. The absence of both sources of sampling error produced a gain of .03 in *MPP* from the use of the best H_d derived 9-test battery over the *a priori* ASVAB.

Table 7

Experiment 1: Average Mean Predicted Performance across 30 Cross samples reflecting Mean Shrinkage due to Sampling Error in Regression Weights and Test Selection[a]

	Best 9-test battery for 9 job families selected using H_d		ASVAB battery
	Test selection error present [b]	Test selection error removed [c]	No test selection
Regression error present [b]	.239 (.042)	.275 (.042)	.243 (.035)
Regression error removed [c]	.291 (.039)	.308 (.035)	.281 (.036)

Notes. All conditions are full least squares estimates.

[a] Standard deviations for the 30 cross-samples are given in parentheses.

[b] Implies use of analysis sample for test selection or assignment regression weights.

[c] Implies use of population for test selection or assignment regression weights.

The above analysis showed the superiority of the ASVAB in the presence of sampling error. This finding resulted from the use of comparatively small analysis samples. This finding, however, was not obtained when a very large analysis sample was utilized. The effects of sampling error, introduced by test selection and by the computation of regression weights using the Experiment 1 analysis was sufficient to affect the outcome of experimental comparisons. This finding has implications for the sample size of back samples used as the source of tests via test selection and regression weights for use in cross validation samples. The design of Experiment 2 permits further analysis of the effect of analysis sample size. The sample sizes chosen in this study were also utilized in previous research studies and thought to be minimum sizes required to make operational decisions.

Experiment 2

Table 8 shows the classification effects of the Experiment 2 simulations for the three different sizes of analysis sample. The comparison between analysis samples formed a between-group factor while all other experimental conditions were repeated measures for 30 different simulated applicant groups.

Comparing across the samples in Table 8, statistically significant classification gains were evident with an increase from N=400 to N=900 and from N=900 to N=1600. The greatest gains from N=400 to N=900 (up to .04 standard scores) were observed for the best three-, five- and nine-test aggregated full least squares (FLS) composites selected from the 29 tests ($F_{1,58}$=15.00, $p<.0001$). Gains and reversals for the same composites from the non-aggregated LSE weights were not significant at $p<.01$. For composites selected from the ASVAB, differences resulting from increasing N from 400 to 900 were not significant at $p<.01$ either for the aggregated LSE job family composites ($F_{1,58}$=4.34) or for the non-aggregated FLS composites ($F_{1,58}$=.0001). The unit weight conditions from N=400 to 900 provided mixed results with no significant gains but one significant reduction for the nine-test composite from the 29-test pools ($F_{1,58}$=10.30, $p<.001$).

Increasing the sample size from 900 to 1600 provided one significant gain in *MPP* among FLS composites ($F_{1,58}$=6.66, $p<.001$ for the five-test aggregate ASVAB composite) and significant gains for all unit weighted composites ($F_{4,232}$=12.21, $p<.0001$ across all samples).

The sample size effect was particularly prominent when comparing the conditions within the separated repeated measure designs within each sample size. Some relationships which were inconsistent at the smaller sample size became more pronounced at both the larger sample sizes. For example, the increase in *MPP* from a three-test to a five-test aggregate LSEs composite became greater in terms of absolute *MPP*, and consistently significant statistically at $p<.0001$, with the increasing sample size .007 in *MPP* ($t_{1,29}$=3.78) to .01 ($t_{1,29}$=3.53) to .04 ($t_{1,29}$=11.45)). Similarly, the effect of adding tests to the unit weighted composites derived from the 29-test pool was inconsistent at N=400 with a gain from three to five tests ($t_{1,29}$=5.25, $p<.0001$) but a decrease from five to nine ($t_{1,29}$=7.60, $p<.0001$). At N=900, this progress showed a decline from three to five tests, albeit not significant ($t_{1,29}$=2.236) and significant from five to nine tests ($t_{1,29}$=13.77, $p<.0001$). At N=1600, the decline in *MPP* was significant for

Table 8

Experiment 2: Average Mean Predicted Performance for Different Analysis Sample Sizes

	PV test selection from ASVAB			PV test selection from Project A 29-test pool		
	3 tests	5 tests	9 tests	3 tests	5 tests	9 tests
Analysis sample N=400						
Aggregate job family test weights	.2304 (.038)[a]	.2378 (.038)	.2288 (.038)	.2059 (.039)	.2468 (.041)	.2957 (.039)
Non-aggregate job family test weights	.2059 (.037)	.2485 (.041)	.2360 (.037)	.2367 (.036)	.2884 (.037)	.3423 (.037)
Unit test weights	.1477 (.041)	.1291 (.041)	-	.1687 (.042)	.1840 (.038)	.1597 (.043)
Analysis sample N=900						
Aggregate job family test weights	.2069	.2163	.2176	.2511	.2791	.3278
Non-aggregate job family test weights	.2111	.2382	.2347	.2359	.2906	.3407
Unit test weights	.1388	.1263	-	.1719	.1662	.1265
Analysis sample N=1600						
Aggregate job family test weights	.2038	.2396	.2344	.2275	.2854	.3361
Non-aggregate job family test weights	.2088	.2431	.2343	.2391	.2854	.3471
Unit test weights	.1699	.1536	-	.2195	.2041	.1681

Note. [a] Parenthetic entries are standard deviations for the 30 cross-samples used for all assignment simulations for N=400.

both three to five tests ($t_{1,29}$=7.60, p<.0001) and for five to nine tests ($t_{1,29}$=5.37). It therefore appears that some relationships were evident at all sample sizes. In general, as sample size was increased, greater consistency across samples, as well as classification gain, was introduced. It is clear that it is advantageous to have 3-test, rather than 5-test, composites when only unit weights are used.

Gains in *MPP* from the addition of two more tests to the three-test composites occurred only in the full least squares weighting conditions of each sample ($F_{1,87}$=1828.19, p<.0001) and this gain increased as the sample size increased ($F_{2,87}$=21.12, p<.0001). The largest gains (.06 in *MPP*) were observed consistently for the 29-test pool aggregate conditions ($F_{1,87}$=736.96, p<.0001) and 29-test pool non-aggregate conditions ($F_{1,87}$=1416.02) across all samples. Further improvements in *MPP* from five-test to nine-test composites also occurred only with the LSEs conditions from the 29-test pool, where there was an average gain of .05 in *MPP* ($F_{1,87}$=2843.4, p<.0001) with no differences across samples ($F_{2,87}$=4.18). As in Experiment 1, there was no apparent asymptote at nine tests in any sample. Earlier studies have demonstrated the potential for further gains by adding more tests to the FLS composite obtained from the 29-test pool. By contrast, all five- and nine-test composites selected from the ASVAB showed no significant differences at p<.01.

In direct contrast to the least squares estimates, the addition of tests to the unit weighted composites, whether from the ASVAB or the 29-test set, resulted in a decrease in classification efficiency. This was the case for all ASVAB five-test unit-weighted composites relative to three-test composites across all samples ($F_{1,87}$=45.02, p<.0001) and all best selected nine-test composites relative to the best five-test composites from the 29-test battery ($F_{1,87}$=353.06, p<.0001). In both cases, the comparison between the best three-test and five-test composites selected from the 29-test battery produced an increase in the smallest sample size ($F_{1,29}$=27.56) but a decrease in the largest sample ($F_{1,29}$=21.07, p<.0001). The reduction from five-test to nine-test composites stayed consistent across samples ($F_{1,87}$=353.10, p<.0001) although the reduction became greater as the sample size increased ($F_{2,87}$=6.62, p<.01).

One explanation for the distinctly different results of increasing the number of tests when different weighting approaches are used is provided by careful consideration of Brogden's *MPP*, that is equal to $f(m)\ R\ \sqrt{(1-r)}^{1/2}$, where R is the mean predictive validity of the tailored (weighted) composites, r is the mean intercorrelation among the tailored predicted performance composites, and $f(m)$ represents an order function according to the number of jobs, m. Brogden (1959) showed that while there is an increase in R, and hence *MPP*, as more tests are added to the composite, this value quickly reaches an asymptote. On the other hand, as r continues to increase the value of *MPP* is reduced. Brogden showed that a large intercorrelation among predicted performance estimates need not imply a trivial classification efficiency if the composites are FLS estimates. However, the magnitude of r continues to effect classification efficiency. In the largest analysis sample of the present experiment (N=1600), r increased in the unit weighted conditions with the increase of tests from three to five using the ASVAB (from .33 to .34) and going from three to nine tests, using the 29-test pool (from .25 to .28). The average validity stayed the same in the case of the ASVAB and increased by .01 from .53 to .54 for the 29 tests. Thus, the unit

weights in particular may be inadequate, in comparison to the FLS estimates, for counteracting the effects of larger assignment variable intercorrelations as the size of the composite is increased.

The unit weighted conditions, overall, provided significantly less *MPP* than both sets of least squares regression weight conditions. Across all sample sizes, a loss from the use of unit weighted composites rather than LSEs resulted both for ASVAB test selection ($F_{1,87}$=3678.99, p<.0001) and for selection from the 29-test pool ($F_{1,87}$=6512.56, p<.0001). The ASVAB unit weights could be improved with the introduction of test selection for three- and five-test-composites from the larger 29-test pool with gains of up to .07 standard scores ($F_{1,87}$=658.95, p<.0001). However, the use of LSEs rather than unit weighted composites exceeded this gain in all cases, and in some, more than doubled the *MPP*s obtained from unit weighted composites.

Differences between LSEs computed from aggregated and non-aggregated job family analysis samples were found in the smaller sample (N=400) ($F_{1,29}$=216.44, p<.0001) but disappeared in the largest sample (N=1600). For the smaller sample, the non-aggregation of job family samples for the computation of weights resulted in higher estimates of *MPP* (up to .05 standard scores). Only the 3-test ASVAB composite reversed this finding when N=400 ($F_{1,29}$=76.91, p<.0001) as well as the 3-test composite from the 29-test pool when N=900 ($F_{1,29}$=24.11, p<.0001). The convergence of these two methods is explained by the increase in *MPP* from the aggregate conditions as the analysis sample size increased. With small sample sizes, non-aggregated job family covariances provide the most stable estimate of the regression parameters; at larger sample sizes, there should be no differences between the two methods. However, data from aggregated samples may be preferred as a better estimate of the population intercorrelations, and would provide, therefore, greater accuracy for experiments in which knowledge of the population parameters is assumed. Also, such differential validity indices as H_d can only be obtained from aggregated samples.

Finally, the selection of tests for composites from an expanded pool of tests compared to the nine ASVAB resulted in consistently higher *MPP*s across all samples for LSE composites ($F_{1,119}$=829.11, p<.0001). In the smallest sample, only the three-test aggregate LSE ASVAB composite resulted in a decrease from the use of 29 tests ($F_{1,29}$=81.33, p<.0001). The aggregate and non-aggregate least squares weighting conditions benefitted equally from the expanded operational battery. However, the nine test composite in particular showed the maximum gain from use of the expanded pool. In the case of the FLS composites, increases of between .07 in *MPP* in the small sample and .11 in the largest sample were obtained by using the best nine-test full least squares composites, optimally selected from the 29 tests, rather than the nine tests of the ASVAB ($F_{1,87}$=3407.02, p<.0001). The difference among sample sizes was also significant ($F_{2,87}$=16.63, p<.0001) indicating a significant increase in the margin of improvement with larger sample sizes. Thus, the improvement over the nine-test ASVAB composite achieved by test selection from the 29-test pool was by far the largest classification gain obtained in this experiment.

Discussion

The two experiments reported here provide evidence that optimal classification provides up to twice as much gain in mean predicted performance as from selection alone. The findings hold direct relevance for both theoretical and practical issues concerning test selection and weight stabilization for assignment composites. In addition, the research provides clarification of the effects of sampling error from test selection and the computation of weights on back samples, and of the effect of back sample size on classification efficiency.

Many of the key findings can be summarized in terms of the contrasting predictions of DAT and the intersection of *g* theory and validity generalization. Figure 4 provides a comparison of these two positions with regard to the design of assignment variables. The first two points of comparison relate to the appropriate method for selecting predictors, either for a single operational battery or separately for job family composites. Selection methods focus only on gains in predictive validity as a measure of the benefit of a test to an assignment composite. Evidence from previous studies has shown that a measure of differential validity, such as Horst's (1954) index of differential efficiency (H_d), provides greater *MPP* when predictors are selected for inclusion in an operational classification battery. Experiment 1 showed, however, that when predictors are selected separately for job family assignment composites, predictive validity is superior H_dm modified to select separately for job families (see Table 5). Further analysis in Experiment 1 suggested that, unless an improved DV index can be formulated to form composites in a way that better replicates the original concepts of Horst (i.e., ensuring overlap of tests across job families and the contribution of differential weights), predictive validity provides the best approach for maximizing the classification efficiency of tailored job family composites selected from a previously selected battery (see Table 6).

The choice of selecting predictors for a single battery or directly for tailored composites is the third issue in Figure 4. The results of Experiment 1 indicated that the gain in classification efficiency over chance assignment resulting from the use of a set of tests selected separately for each job family can be considerable when compared to the use of a single set of tests selected to constitute an operational battery when the two alternative strategies are matched on size of AVs rather than on battery size. A finding of practical significance, for example, was that the selection of a three-test composite by predictive validity provided a gain of .071 standard scores over a single three-test battery selected from the 29 tests (see Table 5). Similarly, the five-test composite improved by .074 in *MPP* over the best single five-test battery and the nine-test composite improved by .061 over the best single nine-test battery.

A similar pattern emerged for the selection of three- and five-test composites from the ASVAB. The gain in *MPP* from separately selecting the tests using predictive validity in each three-test ASVAB composite was .087 over the selection of a three-test battery combined with use of 3-test composites. Equivalent or greater gains could be achieved by increasing the size of the single operational battery from three to five (.073 *MPP* gain) or nine (.110 *MPP* gain). The use of both techniques (i.e., separate test selection for composites and the increase in composite size to five) provided similar gains of .101 in *MPP*

Figure 4
Differences between Selection and Classification Approaches to the Design of Composites

Theoretical/Practical Issues		VG with *g*[a]	DAT
1.	Selection of predictors for an operational battery	Maximize predictive validity	Maximize differential validity for classification and predictive validity for selection
2.	Selection of predictors for tailored job family composites	Maximize predictive validity	Maximize predictive validity
3.	Selection of predictors for batteries versus tailored composites	Two- or three-test composites minimize unstable regression weights	Least squares estimates of the criterion used for tailored assignment composites, but DV used to select batteries
4.	The test pool from which predictors should be selected	Pool most often consists of cognitive ability type predictors (e.g., ASVAB)	Pool should consist of cognitive and non-cognitive measures (e.g., Project A predictors) and vocational information tests
5.	Best composite size	Three tests are adequate	All tests in the battery preferred; not less than five
6.	Composite test weighting	Unit weights provide little if any loss in predictive validity	Least squares weights modified to be positive are stable and provide adequate and appropriate AV composites
7.	Job family validation samples used to estimate regression weights	Analysis sample should be at least 5,000. Small samples cause sampling error which explains much of the variation between validities across jobs	Non-aggregated job family regression weights provide best estimate of betas for up to N = 1200. Aggregated samples for N_j > 1600. LSEs provide improved and substantial *MPP* for N_j > 300. Small samples cause sampling error in test selection and weights calculation. Larger, but obtainable, samples result in higher *MPP*s

Note. [a] Entries attempt to reflect the position often held by some validity generalization proponents who are also *g* theorists.

over a single three-test battery. This gain is provided by a simultaneous increase in both composite and implied battery size. Comparable gains (.096 *MPP*) could only be achieved by the use of H_d as the selection index for the five-test battery.

The advantages of composites over a fixed operational battery are obtained at a cost of a major increase in the size of the implied operational battery and consequent testing time needed to achieve the tailored composites. This drawback is of greater concern with the use of the expanded pool of 29 tests for selecting predictors. In these experiments, composites created from the ASVAB required more or less the same implied battery (i.e., the ASVAB) as the size of the composites increased from three to five to the maximum, nine. Taking the example of the ASVAB FLS composites in Experiment 2 (N=1600), all nine tests were selected for at least one job family in the three-test composites, resulting in the maximum possible implied battery. Selection from the larger pool of 29 Project A tests caused a sizeable increase in the implied battery as the number of tests in the composite increased. Using the same example, selection from the 29-test pool produced an implied battery of 13 tests to achieve three-test composites, 20 tests to achieve five-test composites and 26 tests to achieve the best nine-test composites. In terms of Brogden's equation for *MPP*, as the implied battery increased, the intercorrelation between assignment variables (r) decreased (from .30 to .29 to .28).

The fourth issue in Figure 4 questions the advantage of an expanded pool of predictors from which to select for composites. Theories which rely on maximum predictive validity, generally also lead to a reliance on measures of general cognitive ability (*g*) as the best selection and assignment variables and hence make no requirement for a more heterogeneous group of experimental tests. Experiment 1 (which used some very small job family samples) indicated that the only substantial gain from using the larger battery as the source of composites occurred with nine-test composites (.042 *MPP* gain). The same *MPP* was obtainable for a three-test composite whether the tests were selected from the ASVAB or the 29-tests. However, the gains obtained by adding additional tests to the composite were greater when test selection was from the larger battery. In Experiment 2, the 29-test battery exceeded the performance of the ASVAB in most conditions, and by as much as .11 in *MPP* for the nine-test composites. As the size of the analysis sample increased, gains of .05 and .06 in *MPP* for selection from the 29-test set also became evident for the three-test and five-test composites.

The fifth issue highlighted in Figure 4 concerns the effect of adding more tests to the composite. *g* theorists and validity generalization proponents have promoted only the use of tests in composites which maximize predictive validity. The practice has led to dependence on a single measure of general cognitive ability (*g*), which provides the greatest predictive validity, and perhaps one or two additional measures (such as psychomotor or perceptual tests) which provide some incremental predictive validity. Both present experiments indicate that, while predictive validity is the appropriate measure for selecting tests for tailored composites, the effect of adding more tests to the composites is to increase classification efficiency when tests are optimally weighted. This effect did not appear to level off any faster than the effect of adding more tests to an operational battery (i.e., somewhere beyond nine tests) when

selection was from the 29-test battery. Only when unit weights were used did *MPP* decrease as the composite size increased. It is also of some theoretical interest that the effect on *MPP* of adding to the number of tests directly selected for inclusion in each composite was quite similar to the effect of adding additional tests to a battery when a PV index is used in both cases.

In Experiment 1, increasing test composite size from three to five provided an average increase (per unit increment in the number of tests across all conditions using the 29-test pool) of .013 in *MPP* standard scores. Comparing this gain to the average gain of .012 when the composite size was increased from five to nine, it appears that there was no levelling off of the *MPP* gain as the number of tests reached nine. The selection of tests from the ASVAB, on the other hand, did show a levelling off with the addition of more tests. When composite size was increased from three to five, there was an average increase in *MPP* of .007 compared to .002 from five to nine tests.

In Experiment 2, a consistent increase from three- to five-test composites and from five- to nine-test composites was observed for the 29 tests when the assignment variables were full least squares (FLS) composites. Selection from the ASVAB showed less consistent gain and the use of unit weights completely reversed these findings to produce an average decrease (per increment in number of tests) of .015. Taken together, these results suggest that when assignment composites are least squares estimates, the gain from additional tests continues until somewhere beyond nine tests when predictor selection is from the 29-test Project A pool. When unit weights are used for test composites, the effect of adding tests results in an increase in the intercorrelation of assignment variables (r) and the consequent reduction of *MPP*.

The sixth issue in Figure 4 addresses the appropriate weights for assignment variables. DAT echoes Brogden by stating that FLS composites are optimal in the back sample for the accomplishment of selection and classification. One current popular approach reflected in the Army's system has been the use of unit weighted composites with only three tests in each composite for the purposes of simplification and stabilizing the effects of sampling error on least squares weights and, in turn, predictive validity.

The findings of both experiments reported here indicated a gain from the use of FLS composites with both positive and negative weights permitted compared to alternative weight stabilization methods, and especially unit weighting, in Experiment 2. Experiment 1 also indicated that other approaches to weight stabilization maintained a high level of classification efficiency. In particular, the restriction of weights to positive at the test selection stage resulted in *MPP*s comparable to the best weighted composites selected from both the ASVAB and the 29-test pool. Indeed, the loss resulting from the implementation of this restriction was less than (about one-third) the loss from decreasing the number of tests in the composite from five to three and permitting negative weights. Thus the introduction of a positive-weights constraint during test selection for FLS composites provides both stability of regression estimates and gains in classification efficiency.

Issue 7 of Figure 4 addresses the size of analysis samples required to conduct classification research. Experiment 2 indicated that utilizing the non-

aggregated data for separate job sample regression weights provided stable estimates of weights and increased classification efficiency at the smallest sample size. As the analysis sample increased in size approaching the population (i.e., N=1600), the aggregate and non-aggregate methods converged to no difference. Since at larger sample sizes, therefore, aggregate sample data provides an improved estimate of population intercorrelations of predictors. The final issue in Figure 4 differentiates the selection and classification positions as to the appropriate sizes of back samples for test selection and calculating weights. Most theorists use sampling error resulting from small empirical validation samples as the major explanatory factor for variation across predictor validities, leading to the position that once sampling error is controlled along with other artifacts, validity generalization is more possible. The DAT position also recognizes the effects of sampling error in back samples on test selection and the calculation of regression weights. However, this is used to explain reductions in classification efficiency rather than focusing on predictive validity. The results of Experiment 2 indicated that classification efficiency increased as the size of the analysis sample increased and that the results of comparisons between other factors became more consistent. Larger sample sizes apparently provide better estimates of population intercorrelations and validities and, therefore, account for more accurate test selection and more stable regression weights.

Theoretical Conclusions and Operational Implications

A major purpose of the study was to examine the usefulness of differential assignment theory (DAT) in the resolution of practical issues that arise in reconstituting aptitude area composites. Several DAT principles keyed to the reconstitution objective were examined in the combined context of theoretical expectations and study results. A reassessment of several DAT principles relevant to the design of future classification systems and to the adjustment of research methodology for future reconstitution procedures is provided in the context of the results of this study.

DAT principles, believed to be sufficiently well-established to be applicable to the construction of new and improved AA composites include the following:

1. The best test composites for either selection or classification are least squares composites (LSEs), although composites based on factors transformed into simple structure can do almost as well if the job families are also in simple structure (Statman, 1993).

2. An increase in battery size provides a steady increase in classification efficiency, as measured by *MPP*, and a critical point where a further increase in the number of tests would provide only a trivial increase in *MPP* has not been established.

   a. The relationships between *MPP* and number of tests in test pools, operational type batteries of tests, and aptitude area type test composites have important differences. Previous research established a positive relationship between number of tests in batteries and *MPP* that

continued up to n = 29; prior to this study no data were available regarding the effect of composite size on *MPP*.

3. Brogden's 1959 model of *MPP* provides an approximation of the relationships of the validities (*R*) of LSEs, the intercorrelation (*r*) among these LSEs, and *MPP*; an increase in *MPP* will clearly result from an increase in *R* and a decrease in *r*.

a. The value of *R* has a positive relationship to the number of tests (n) in the AA composite, but the value of *R* rapidly approaches the limit that can be achieved by selecting from a given test pool as n is increased.

b. The greatest hope for increasing *MPP* from either test selection or from the reclustering of jobs into job families is in the obtaining of a smaller value of *r*.

c. A smaller value of *r* tends to result from an increase in battery size, while the effect of an increase in composite size when tests are selected from a fixed size battery is more likely to increase the value of *r*, the use of unit weights instead of best weights also increases the value of *r*.

4. Statistical theory and the results of many empirical studies lead to the expectation that the sampling error in a validity coefficient has an approximately linear relationship to the square root of the size of the average analysis sample for each job family.

a. The sampling error in *MPP* includes an aggregation of the sampling errors of each assignment variable in each job family analysis sample. The inclusion of several small job family samples will increase shrinkage and reduce the stability of the *MPP* estimate more than will be compensated for by an equal number of larger job family samples (on a scale defined by the square root of N).

b. The nine job family analysis sample sizes of Experiment 1 range from the four smallest samples of 129, 203, 289, and 464 to five above average sized samples which together yield a mean N of 775, a combination of N_j that yields more shrinkage and less stability than the smallest set of equal sized N_j (N_j = 400) used in Experiment 2.

5. Differential validity indices have proven superior to predictive validity indices for sequential selection of tests for batteries where the objective is to increase *MPP* obtainable from the use of FLS composites in making optimal assignments to jobs: New research was required to determine whether modified DV indices would provide a similar superiority for selecting tests to be included in test composites for each job family.

The first of the DAT principles listed above has been consistently confirmed during the past five years in every model sampling experiment involving DAT conducted at George Washington University by Zeidner, Johnson and colleagues. Comparisons were made between LSEs and LSEs modified to eliminate negative

weights, or between LSEs and LSEs computed after all variables yielding negative weights were avoided in the selection process. Further comparisons were made between LSEs and unit weighted composites. Although a wide difference between *MPP* provided by LSEs and by the use of unit weights was found, the unit weighted composites based on test selection provided a sizable gain over the *MPP* provided by a priori composites.

The second of the DAT principles has also been consistently confirmed for batteries but do not appear to be unconditionally true for test composites used as AVs. When the test pool is large enough so that the implied battery does not quickly encompass the entire pool, it is the implied battery size rather than the composite size that predicts *MPP*. On the other hand, when the implied battery size and composite size become the same, as in the case of the ASVAB condition in this study, *MPP* continues to increase slowly as tests are added to the composites.

Under the ASVAB conditions of this study, even a 3-test composite would include all the tests in the pool in at least one composite. An increase in *MPP* is obtained as the composite size increases, but this gain is comparatively small. It appears to be necessary to increase battery size in order to achieve a major increase in *MPP*. Unfortunately, an increase in battery size is both financially and administratively costly.

The third principle relies primarily on evidence from other studies (e.g. Statman, 1993), but is consistent with results of this study. The importance of *r* in the prediction of *MPP* cannot be minimized. Methods intended as means of reducing *r* are always worthy of consideration in the reconstitution of either AA composites or job families.

The fourth principle is also confirmed but the investigators, if repeating this study, would use N_j = 1600 for all job families, instead of the Project A sample sizes. While we are interested in seeing how much the shrinkage phenomenon affects *MPP* in the smaller and irregular sample sizes of Experiment 1, we would have liked to have had a more stable depiction of the relationships among the classification related variables than resulted from the use of such small samples for some job families.

Note that the use of 9-test LSEs from the ASVAB as AVs provides as much *MPP* as does the best 9-tests selected from the 29-test pool and used as LSE composites. It is clear that the job family sample sizes used in Experiment 1 are not sufficiently large to justify changes in the ASVAB based on test selection from the 29-test pool--regardless of which index is used. Much of this weakness is due to the heterogeneity of the particular jobs found in the Project A concurrent study, i.e., some of the combined jobs appear to provide as much heterogeneity within job families as across job families.

The question as to whether the last of these principles, the superiority of a DV index for selecting tests for batteries, can be extended to the selection of tests for composites was answered with respect to $H_d m$. The principle still holds with respect to assignment to jobs (as contrasted to the representations of job families provided by the Project A data) and when test selection is also for batteries (with assignment by FLS composites), instead of separately to AVs.

Horst's index of differential validity (DV) cannot be used to select tests separately for each job family composite without major modifications that would result in an entirely different index.

It seems reasonable to assume that a different test selection method than any of those investigated in this study--one which minimizes test overlap in the composites while minimizing the loss in predictive validity--would be superior to the PV index for the direct selection of tests for composites. Another alternative DV index could be based on the elimination of "Brogden *g*" from the *R* and **V** matrices of the analysis sample, providing a means of test selection that would not be unduly effected by the presence of a non-productive *g* component in some tests. The elimination of this component would reduce overlap in composites and could not, except by chance, reduce *R* to an extent that is not offset by the reduction in *r*. This issue with respect to composites was not resolved in the present study.

The relationship of *MPP* to a number of conditions not previously investigated are explored in this study. These conditions include the effective use that unit weights and positive only weights has on the *MPP* provided by sets of AVs. The AVs modified to have positive weights were otherwise close approximations of LSEs, since the negative weights were usually small (and few in number), the tests deleted in order to assure positive weights would not have made other than small contributions.

While it was not predicted that unit weighted AVs would provide strong competition to LSEs for use as AVs, the use of unit weighted composites demonstrates the contribution that can be provided by test selection, even in the absence of least squares weighting. It would appear that unit weighted composites should contain more than three tests, and, possibly, some unit weighted composites should not exceed two tests to achieve maximum potential classification efficiency.

From a theoretical point of view, the steady and early decrease in *MPP* as the number of unit weighted tests in a composite is increased demonstrates how important *r* is as compared to *R* in unit weighted composites. These results show that the number of overlapping tests in a set of unit weighted test composites is very important with respect to the potential classification efficiency (*MPP*) of the system.

An analysis of existing data to provide a proposed reconstitution of the Army aptitude areas is expected to begin in the fall of 1993. A number of the issues investigated in this study pertain to how samples and sets of predictor variables should be selected for analysis, and how tests should be selected for job family composites. However, the results of this study have other operational implications that concern future, longer range, applied studies for the improvement of the operational classification system.

It is important to pay particular attention to the size of the smaller validity samples used to compute AVs. Serious consideration should be given to the job families for which the combined analysis sample sizes are less than 1000 (but $N_j > 2000$ is highly desirable). Additional jobs should be added to achieve

this minimum size, or job families should be combined. Large job families should be considered for shredding into two or more homogeneous job families in order to provide a better quality distribution as well as to provide a higher overall *MPP*.

The higher level of *MPP* provided by assignment to jobs, as compared to assignment to the operational job families (as represented by the jobs in the Project A concurrent study), leads to the conclusion that some of the LSEs are not appropriate for assignment to all of the jobs contained in one job family. On this basis, the importance of choosing a set of jobs that can accurately represent each job family in a particular analysis intended to identify the LSEs for each job family should be emphasized.

The PV index should be used as the primary test selection index for developing test composites until, and if, further research can establish the effectiveness of a DV index which can reduce test overlap (i.e., reduce *r*) while minimizing a decrease in *R*. Any test index used to select tests for operational composites should be modified to reject tests yielding negative weights in a LSE. The *MPP* provided by use of the PV index is impressive and can be expected to provide a major improvement in classification effectiveness, even if the further gain that might be provided by use of an improved DV index is not immediately available.

The test selection index with the greatest potential for showing a superiority over the use of *R* alone is believed to be the product of **R** and the square root of $(1 - r)$. The use of this index as the figure of merit for selecting tests could be optimally implemented by selecting the best combination of tests, rather than through the use of a sequential algorithm based on either accretion or deletion. The process of seeking optimal combinations of tests for a figure of merit that includes *r* will require an iterative estimation of the other 8 AVs while computing an average *r* to permit the finding of the best interim AV for each job family.

The non-aggregated analysis samples should be utilized for both test selection and computation of weights for operational test composites. Aggregated analysis samples remain appropriate for use as a designated population and as the source of DV indices for the development of batteries. An adequate sized N is especially important for research situations where an aggregated analysis sample is to be used.

Appendix A: Criterion Issues Important to the Conduct of Selection and Classification Research

This appendix expands explanations of criterion issues partially explained in the main body of the report and introduces additional criterion related issues. The discussion of these issues are intended to assist in the interpretation of our experimental results in the context of DAT--both in this and previous research. Clarification of the issues treated here have considerable importance as to how future DAT research should be conducted and to the design of improved selection and classification systems incorporating DAT principles.

We discuss criterion issues in four categories: (1) differences in the core technical proficiency (CTP) criterion between Batch A and Z jobs--when optimal assignment is to a set of jobs that includes jobs from both batches; (2) choosing CTP from among five available criterion components for the conduct of classification research, although all five are considered to be important for the conduct of selection research; (3) choice of a criterion variable when comparing the efficacy of one-stage and two-stage selection and classification strategies; and, (4) the effect of grouping either predictor or criterion scores into intervals.

1. Batch A. vs. Batch Z Criteria

The CTP criterion component can be divided into a "hands on" subcomponent and a carefully crafted job information subcomponent; both subcomponents being job specific. The CTP criterion for Batch A jobs consist of a combination of both subcomponents while the CTP criterion component used in Batch Z jobs does not contain a "hands on" subcomponent. Both in the present research and in several previous research efforts, one of the 19 MOS for which criterion and predictor data was available from the concurrent study of Project A was eliminated because of a very small sample size.

In the previous research the remaining 18 jobs were equally divided into Batch A and Z sets. Thus optimal assignment to no more than nine jobs could be accomplished without using predicted performance based on differing types of criterion variables in the same assignment process. One question that can be raised concerns the effect of using different types of criterion variables in the same assignment and evaluation process on classification efficiency results. The 9 MOS in each batch were considered separately and together in several model sampling experiments comparing the efficacy of differential validity (DV) and predictive validity (PV) indices for selecting tests to be used in classification test batteries, and for comparing the *MPPs* provided by optimal assignment to 9 and 18 jobs (Johnson, Zeidner, & Scholarios, 1990). The gain in MPP resulting from using the DV index compared to the PV index was .048 in the 9 Batch A jobs as compared to .036 in the 9 Batch Z jobs, and -.008 in the combined set of 18 jobs. The heterogeneity of job criteria definitely reduces the sensitivity for detecting a superiority of DV compared to PV indices for use in selecting tests for batteries.

There is a surprising reduction in the gain in MPP, one not predicted from DAT, resulting from the substitution of DV for PV indices--when all 18, instead of 9, jobs are utilized as surrogates for job families in the assignment process. This reduction might be at least partially explained in terms of the heterogeneity of the criterion variables across the two halves of the 18 job set. However, the major reason for using all 18 MOS together was to investigate the gain in MPP resulting from increasing the number of assignment targets (surrogates for job families) from 9 to 18. The amount of gain from using 18 assignment targets instead of 9 jobs was approximately what we expected from a consideration of Brogden's 1959 model and DAT. Better but comparable results were obtained by Johnson, Zeidner, and Leaman (1992) where differing numbers of job families could be compared in a model sampling experiment using SQT as the criterion variable for all MOS and job families. Obviously, the criterion heterogeneity across the batch A and Z jobs did not interfere with DAT predictions regarding the effect of increasing the number of job families had on classification efficiency.

Since the present study makes optimal assignments to job families, rather than to MOS, any effect of criterion heterogeneity is greatly diluted. Our use of the concurrent data results in most job families being represented by one Batch A MOS and one Batch Z MOS. However, one Batch Z family (i.e., electronics) contained only one MOS. In this latter case, sampling error due to the small size of the validity sample (N = 123) was of more concern than was the lack of a "hands on" criterion subcomponent in the job family criterion.

2. Selecting a Criterion Component for Use in Classification Research

Ideally the criterion variable is a measure of the capabilities good performers possess to a greater extent than do poor performers. Thus, the value management places on employees in a given job is believed to be highly correlated with the amount of these capabilities displayed on the job. Some of these capabilities, particularly in the Army and Marines, are required to a similar degree in all jobs. The five Project A criterion components contain only one which could be reasonably specific to a designated MOS.

It would appear that the designers of the Project A criterion components deliberately constructed only one of these components (i.e., CTP) to be job specific. For example, it would have been possible to construct a job specific leadership subcomponent for each MOS for which CTP was constructed. Instead, the project A criterion designers defined and developed a general Army wide leadership component. A job specific leadership component would have emphasized managerial capabilities for some MOS, ability to provide technical instruction and guidance to a technical process in some MOS, being a role model in various job specific ways in others, and, in still others, inspire subordinates to follow the indicated individual into dangerous situations.

While all five of the Project A criterion components are clearly appropriate for use as criterion variables for selection research, only one appears appropriate for classification research. The empirical research reported by Wise et al. (1990) supports this conclusion. The inclusion of these other four criterion components in the research criterion used in a developmental process for a classification system would, at best, have the same effect as the

addition of either sampling error or otherwise irrelevant variance. An increased sample size can compensate for an increase in unbiased error variance during research.

The most intuitively obvious effect of adding components to a criterion composite that are effective for selection, but not for classification, for use in developing a personnel classification system is the addition of variables which primarily measure Brogden's *g*, predictors which are equally valid across all job families. The other four Project A criterion components appear to be heavily loaded with Brogden's *g*. Brogden theorized and provided credible proofs that the addition of such variables had little or no effect on classification efficiency (Brogden, 1959, 1964).

If the addition of components containing little or no relevant classification variance has no bias other than an excessive loading on *g*, the dilution of the composite criterion through the use of these components will not bias decisions made regarding test selection or computation of weights for tests in AVs. This particular lack of bias is because the same test battery can approach the maximum for both selection and classification: the increase of selection efficiency does not decrease classification efficiency if this done by increasing the amount of Brogden *g* in the tests included in the battery. However, the introduction of Brogden *g* into the criterion variable will provide an unfortunate bias with respect to most other system design choices that affect classification efficiency in the design of a personnel classification system.

Even if there was no bias introduced through the inclusion of a criterion component that is irrelevant to classification, it is certain that this inclusion will reduce the sensitivity of a composite criterion regarding classification system design decisions. The use of a larger sample size can replace this sensitivity, if, and only if, the addition of Brogden's *g* to the joint predictor-criterion space was the only effect of adding the irrelevant components.

Very biasing effects of including irrelevant components in a criterion composite can be expected when the criterion is playing the role of an evaluation variable, that is, for obtaining the test weights to be applied to predictor scores to create an evaluation variable. While the effect of using a composite criterion instead of CTP in the process of obtaining AVs may be easily counteracted by increasing the size of the Ns in the validation samples, a quite different result can be expected from including the irrelevant components in the evaluation criterion. Thus inappropriate decisions regarding system design and strategies would result from the use of these irrelevant components as a part of the evaluation criterion variables.

One can subject the effect of using the different possible criterion variables in the analysis sample to a scientific test with respect to an agreed upon criterion variable used in the evaluation process. If the agreed upon evaluation criterion is CTP, we believe the use of the composite criterion variable used in conjunction with appropriately larger sized Ns to form AVs would provide essentially the same sized *MPP*s as when CTP is used to form AVs. In contrast, the use of a composite containing the four Project A criterion components other than CTP as the evaluation variable would not provide an MPP significantly different from zero for any method of forming AVs.

Consider a criterion component for possible addition to an already existing criterion composite. Assume the new component under consideration measures an attribute highly important to job performance that is not adequately measured by the existing criterion composite. Also assume that the component, unlike the composite, lacks key psychometric characteristics highly desirable for classification research while both composite and the component possess the psychometric characteristics most essential for selection research.

We will consider two possible roles for the augmented composite's candidate component, for use in: (1) an analysis sample to select tests for batteries or AVs, compute test weights, cluster jobs into families, etc.; and, (2) to compute test weights for evaluation variables (i.e. predicted performance) to provide *MPP* scores for making a variety of system design decisions.

We believe that if the objective is to provide classification efficient test batteries, assignment variables, or both, the use of the augmented composite will provide little or no harm when used in the first role. For example, tests with a higher relationships with *g* would undoubtedly result from substituting the augmented composite for the previous classification efficient criterion composite. DAT predicts that selecting and adding tests with higher *g* loadings would have little effect on classification efficiency other than requiring larger samples sizes to conduct research. Most other system design decisions (other than test selection) made independently of the evaluation process, as in role 1, would be biased by the addition of a criterion component which is inappropriately loaded with *g*.

It is generally true that variance in a criterion component that is irrelevant to classification, will, if other than error variance, be correlated with measures of selection efficiency and other system characteristics that may be in competition with classification efficiency. Thus, when the criterion composite is to perform role 2, critical decisions regarding the design of a classification system will be biased by the inclusion of a criterion component that is relevant for selection , but not for classification.

The appropriateness of a criterion component for use as a measure of the value of a soldiers capability to perform on an Army job can be disputed by showing that the component has undesirable psychometric qualities. However, a criterion component must also measure a performance capability that is judged to be essential to the accomplishment of job duties. Even if the component can be validated against a more ultimate criterion variable, the relevance of that more ultimate criterion has to be established by judgment. However, assuming that a component passes the judgment test, we can still reject a component for use as either a selection or classification criterion if it fails to pass certain psychometric tests.

To be appropriate for selection research, a criterion component should be a reliable measure of capabilities required in either all, or at least a substantial number, of the jobs to which a selected individual might be assigned. In addition to measuring a capability required for job performance at the desired level, a criterion component being considered for use in selection research should be required to pass a psychometric test, a requirement which is comparable to, but different from, the psychometric test we believe should be

applied to criterion components being considered for use in classification research. For example, a criterion component being considered for use in selection research should be scaled in such a way that experimental conditions can be expected to affect component scores at the performance levels considered to be most important for system utility. No matter how important expert judges consider the attribute believed to be measured by the criterion component, if the actual measure is not sensitive to differences in the amount of this attribute possessed by a soldier at the desired range on the measurement scale, the criterion scale should not be included as a part of the criterion variable for selection research.

The use of both selection efficient variables and classification efficient variables, if introduced into a battery through the use of a criterion variable deemed to be suitable for both selection and classification processes, will not measure either the total system potential for either selection nor classification. However the use of a criterion not suitable for selection to make choices concerning selection systems can harm SE, and similarly, the use of a criterion not suitable for making assignments across jobs can harm CE.

In summary, it would appear that the use of a composite criterion that is unbiased except for the heavy loading on *g* of some of its components will do little harm to the construction of AVs to be used in the classification process This harm can be usually remedied by the use of a larger N in the analysis sample. However, a composite criterion that has such a large loading on *g* as to make the component irrelevant to classification, could adversely affect many other kinds of system design decisions, as compared to the use of a criterion variable that is entirely classification relevant. Any selection/classification system bias, other than the inclusion of Brogden *g*, can adversely affect all decisions regarding system features. Even the presence of Brogden *g* constitutes a serious bias with regard to many system design decisions. We believe the best overall selection and classification system can be obtained, using the current state of the art, using separate criterion variables for selection and criterion research in the design of a two-stage selection and classification system. Further basic research, with particular focus on criterion issues, must be successfully accomplished before a one-stage simultaneous selection and classification system can be designed and installed--unless a criterion like CTP or SQT for use in both selection and classification research is acceptable to both researchers and management.

3. Criterion Issues Bearing on the Comparison of Two vs. One stage Strategies

The traditional two-stage strategy calls for selecting from an applicant pool into the organization (e.g., Army) using a single selection variable (SV). The AFQT is presently the SA for all services. However, a better measure of *g* would most likely provide greater selection efficiency (SE). The second stage classifies selected personnel to job families using aptitude areas as assignment variables (AVs), and then assigning to an MOS within the appropriate job family. A one-stage strategy calls for the simultaneous selection, classification, and assignment using the same AVs to effect both selection and assignment to MOS. Although the one-stage strategy is more efficient and equitable under certain assumptions, the two-stage strategy is the one used by all services and is the one simulated in this research study.

A one-stage, simultaneous selection and classification strategy, is clearly superior to a two stage strategy when both stages are evaluated in terms of the same criterion variable as is used to evaluate the one stage strategy (Johnson and Zeidner, 1991; Whetzel, 1991). However, this superiority cannot be expected to hold if the best criterion variable for each of the two stages is used to design and evaluate the corresponding stage (as recommended above).

As stated above, we believe it is usually desirable to utilize a different criterion variable when conducting research on selection of applicants for admittance into an organization, as contrasted to the best criterion variable for use in conducting research on personnel classification. If no classification efficient criterion exists, new employees who have already been accepted into the organization may just as well be randomly assigned, or assigned in accordance with criterion neutral preferences.

When a classification efficient criterion is available, a two-stage strategy will usually be profitable. With such a strategy, selection can be viewed as a preliminary process which can have considerable effect on the classification process; the lower the selection ratio (SR) and/or the greater the selection efficiency is, the higher will be the *MPP* obtained in the second stage that is entirely due to the classification and assignment process. In contrast, a change in the efficiency of the second stage has no effect on the efficiency of the first stage.

4. The Differing Effects of Grouping Scores into Intervals for Selection vs. Classification Research

The use of operational decisions such as: (1) pass vs. fail in schools, (2) non-promotion vs. promotion, or (3) no special recognition vs. special recognition, is occasionally proposed as a more realistic alternative to the use of performance evaluations or job knowledge tests that yield many intervals in the distribution of criterion scores. The use of simulation models which call for using a small number of intervals in either the assignment variables or the evaluation (i.e. utility) variables is also tempting.

It is well known that decreasing the number of intervals on a criterion scale from a moderately large number to two, thus creating a dichotomous variable, or to 9 (resulting in stanine scores) will decrease the magnitude of validity coefficients. The differing effect that results from reducing the number of intervals in the upper, or alternatively in the lower, half of a criterion distribution when using SRs less than .5--in personnel selection as compared to classification research--is not so well known. Model sampling research in which the reduction in magnitude of *MPP* resulting from the reducing of the number of intervals in the assignment variables in an optimal assignment process are reported by Sorenson (1967).

Sorenson showed that grouping continuously distributed assignment variable scores into 9 intervals has much more effect on classification efficiency than on selection efficiency. He also showed that if the entire range of scores is represented by a small number of intervals, it is better for a classification process to concentrate these intervals in the upper half of the range than in the lower half. This contrasts with the selection process where there is more

advantage in using smaller intervals in the lower half of the range of AV scores. For similar reasons grouping criterion scores into intervals is more damaging to classification efficiency than to selection efficiency. It is essential that the full range of criterion scores be utilized rather than the substitution of pass/fail, satisfactory/unsatisfactory, or similar dichotomous variables when classification research is conducted.

References

Burket, G. R. (1964). A study of reduced rank models for multiple predictors (Research Monograph). Pittsburgh, PA: American Institutes for Research and University of Pittsburgh.

Campbell, J. P. (Ed.). (1987). Improving the selection, classification and utilization of Army enlisted personnel: Annual report, 1985 fiscal year (Tech. Rep. No. 746). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A193 343)

Horst, P. (1954). A technique for the development of differential prediction battery. Psychological Monographs, 68, (9, Whole No. 380).

Horst, P. (1954). A technique for the development of differential prediction battery. Psychological Monographs, 69, (9, Whole No. 380) 1-22.

Horst, P., & MacEwan, C. (1960). Predictor eliminator techniques for determining multiple prediction batteries. Psychological Reports, 7, 19-50.

Hunter, J. E., Crosson, I. J., & Friedman, D. H. (1985). The validity of the Armed Services Vocational Aptitude Battery for civilian and military job performance. Rockville, MD: Research Applications Inc.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological bulletin, 96, 72-98.

Johnson, C. D., & Zeidner, J. (1991). The economic benefits of predicting job performance. Vol. 2: Classification efficiency. New York: Praeger.

Johnson, C. D., & Zeidner, J., & Leaman, J. A. (1992). Improving the classification efficiency by restructuring Army job families (Tech. Rep. No. 947). Alexandria: VA. U.S. Army Research Institute for the Behavioral Sciences. (AD A250 139)

Johnson, C. D., Zeidner, J., & Scholarios, T. M. (1990). Improving the classification efficiency of the Armed Services Vocational Aptitude Battery through the use of alternative test selection indices. (IDA Paper P-2427). Alexandria, VA: Institute for Defense Analyses.

Maier, M. H., & Grafton, F. C. (1981). Aptitude composites for ASVAB 8, 9 and 10 (Research Rep. 1308). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A109 471)

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Project A: Validation of current and alternative ASVAB area composites, based on training and SQT information on FY 1981 and FY 1982 enlisted accessions (Tech. Rep. No. 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A156 807)

Naylor, J. C., & Shine, L. C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. Journal of Industrial Psychology, 3, 33-42.

Schmidt, F. L., Hunter, J. E., & Larson, M. (1988). General cognitive ability versus general and specific aptitudes in the prediction of training performance: Some preliminary findings. Navy Personnel Research and Development Center: San Diego, CA.

Scholarios, T. M. (1990). Maximizing potential classification efficiency: Selection of predictor measures based on alternative psychometric indices. Unpublished doctoral dissertation, The George Washington University, Washington, DC.

Scholarios, T. M., Johnson, C. D., & Zeidner J. (in press). Selecting predictors for maximizing the classification efficiency of a battery. The Journal of Applied Psychology.

Sorenson, R. C. (1967). Amount of assignment information and expected performance of military personnel. (Tech. Research Rep. 1152). U.S. Army Personnel Research Office: Washington, DC.

Statman, M. A. (1992). Developing optimal predictor equations for differential job assignment and vocational counseling. Paper presented at the annual Convention of the American Psychological Association, August 14-18. Washington, DC.

Statman, M. A. (1993). Improving the effectiveness of employment testing through classification: Alternative methods of developing test composites for optimal job assignment and vocational counseling. Unpublished doctoral dissertation, The George Washington University, Washington, DC.

Whetzel, D. L. (1991). Multidimensional screening: Comparison of a single-stage personnel selection/classification process with alternative strategies. Unpublished doctoral dissertation, The George Washington University, Washington, DC.

Wise, L. L., McHenry, J., & Campbell, J. P. (1990). Identifying optimal predictor composites and testing for generalizability across jobs and performance factors. Personnel Psychology, 43, 335-366.

Young, W. Y., Houston, J. S., Harris, J. H., Hoffman, G., & Wise, L. L. (1990). Large-scale predictor validation in Project A: Data collection procedures and data base preparation. Personnel Psychology, 43, 301-312.

Zeidner, J. (1987). The validity of selection and classification procedures for predicting job performance. (IDA Paper P-197). Alexandria, VA: Institute for Defense Analyses.

Zeidner, J., & Johnson, C. D. (1991a). Classification efficiency and systems design. Journal of the Washington Academy of Sciences, 81, 110-128.

Zeidner, J., & Johnson, C. D. (1991b). The economic benefits of predicting job performance. Vol. 1: Selection utility. New York: Praeger.

Zeidner, J., & Johnson, C. D. (1991c). The economic benefits of predicting job performance. Vol. 3: Estimating the gains of alternative policies. New York: Praeger.

Zeidner, J., & Johnson, C. D. (in press). Is personnel classification a concept whose time has passed? In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), Personnel selection and classification: New directions. Hillsdale, NJ: Erlbaum.