Technical Report 1071

The Substitutability of Criteria in the Development and Evaluation of ASVAB Classification Procedures

Joseph Zeidner
The George Washington University

Cecil Johnson
The George Washington University

Yefim Vladimirsky
The George Washington University

September 1997

United States Army Research Institute for the Behavioral and Social Sciences

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

The George Washington University

Technical review by

Frances Grafton, SARU
Peter Legree, OPRRU

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to : U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: TAPC-ARI-PO, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

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

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

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
1997, September	Final	September 1993-July 1996

4. TITLE AND SUBTITLE

The Substitutability of Criteria in the Development and Evaluation of ASVAB Classification Procedures

5a. CONTRACT OR GRANT NUMBER

MDA903-93-K-0014

5b. PROGRAM ELEMENT NUMBER

0603007A

6. AUTHOR(S)

Joseph Zeidner, Cecil Johnson, and Yefim Vladimirsky (George Washington University)

5c. PROJECT NUMBER

A792

5d. TASK NUMBER

1122

5e. WORK UNIT NUMBER

C04

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The George Washington University
Office of Sponsored Research
2121 I St., NW, Suite 601
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: TAPC-ARI-RS
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. MONITOR ACRONYM

ARI

11. MONITOR REPORT NUMBER

Technical Report 1071

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

COR: Peter J. Legree

14. ABSTRACT *(Maximum 200 words)*:

The major goal of this research is to determine the adequacy of using operational skill qualification test (SQT) measures to serve as a criterion surrogate for the more widely accepted, but prohibitively costly, hands-on measures. If it could be shown that similar decisions are made or similar outcomes are obtained in the classification context, the criteria would be considered as substitutable for one another. Project A's longitudinal data set was used to obtain hands-on and job knowledge measures for 9 Military Occupational Specialties (MOS) and only job knowledge measures for an additional 6 MOS, 18 to 24 months after the cohort sample entered the Army. Operational Armed Services Vocational Aptitude Battery (ASVAB) and SQT scores for the same cohorts were available from official records. Findings indicate close similarities in: selecting tests for assignment composites, patterns of predictor test validities; classification efficiency (MPP); and in factor structure in the joint predictor-criterion space. The overall conclusion is that either criteria can serve as a surrogate for the other in developing classification procedures using ASVAB.

15. SUBJECT TERMS ASVAB; Differential Assignment Theory; Army existing aptitude area composites; Overprediction and underprediction of minority criteria performance; Classification and assignment; MOS; Operations Research of personnel selection and assignment; Personnel selection and classification; SQT; Fairness of predictor composites by race and gender; Model sampling experiment; Mean predictive validity

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	45	Michael G. Rumsey (703) 617-8275

Standard Form 298

Technical Report 1071

The Substitutability of Criteria in the Development and Evaluation of ASVAB Classification Procedures

Joseph Zeidner
The George Washington University

Cecil Johnson
The George Washington University

Yefim Vladimirsky
The George Washington University

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

September 1997

Army Project Number
2O363007A792

Manpower and Personnel

FOREWORD

This report is one of a continuing series of research efforts designed to improve the selection and classification efficiency of the Armed Service Vocational Aptitude Battery (ASVAB). In a related research effort, the authors made recommendations for a new classification system that called for replacing the operational aptitude areas with assignment composites based on least squares estimates (LSEs) of the criterion specific to each job family. The present report bears on the credibility of the criterion measure employed, the Skill Qualification Test (SQT), a paper-and-pencil-based operational measure of job knowledge proficiency widely used in the personnel system of the Army.

While hands-on measures of job proficiency are considered by many to be benchmarks of job proficiency, their costs are prohibitively expensive for classification research on 250 entry-level jobs. There is a need to examine a less costly job knowledge measure, the SQT, as the criterion of job proficiency in making decisions about ASVAB use for classification procedures. The overall goal of this study, then, is to determine the adequacy of SQT measures to serve as surrogates for hands-on job proficiency measures. If it could be shown that similar decisions or similar outcomes are obtained for either criterion, then SQT would be considered as substitutable.

Comparisons are made between the two measures dealing with the similarities of decisions reached such as test selected for use in assignment composites. Similarities of outcomes are also judged such as mean predicted performance, reliability, stability of results using other data sets, intercorrelations among criteria, and factor structure in the joint predictor-criterion space.

The overall conclusion reached is that either the hands-on or job knowledge criterion can serve as the surrogate for the other when conducting classification research on the ASVAB since the results indicate that similar decisions are made in the developmental process and outcomes are judged to be equivalent. This is a critically important finding because it enables researchers to continue classification research using a practical paper-based criterion with demonstrated validity, reliability, relevance, and acceptability.

ZITA M. SIMUTIS
Technical Director

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. We are grateful to Peter Legree for his helpful comments on the draft report and also to Frances Grafton for her invaluable assistance in preparing the database and her review of the draft report. The Contracting Officer's Representative for this effort was Dr. Peter Legree.

THE SUBSTITUTABILITY OF CRITERIA IN THE DEVELOPMENT AND EVALUATION OF ASVAB CLASSIFICATION PROCEDURES

EXECUTIVE SUMMARY

Research Requirement:

In a related research effort, the authors proposed a new classification system to replace the aptitude area composites presently used by the Army. The proposed system would use an invisible black-box first tier in which separate least squares estimate (LSE) composites are computed for each of approximately 70 core Military Operational Specialties (MOS) forming families among the about 250 entry-level Army MOS. A visible, revised second tier system, in which 9 to 16 families encompass all MOS, is proposed for recruiting, counseling, and administration. The proposed changes were based on an analysis of the classification efficiency of various assignment composites using operational Skill Qualification Tests (SQT) as measures of job proficiency. However, before full consideration be given to replace the current aptitude composites, the suitability of using SQT as a criterion measure of job proficiency needs to be carefully evaluated.

The overall objective of the present research is to determine the adequacy of SQT measures of job knowledge to serve as surrogates for core technical proficiency (CTP) measures of hands-on job proficiency in developing classification procedures for ASVAB assignment variables. While hands-on performance measures are considered by many to be benchmarks of job proficiency, their costs in construction and administration make them prohibitively expensive for classification research on core job families to represent more than 20 entry-level jobs. The SQT would be considered an adequate substitute for CTP if it could be shown that the same developmental decisions are reached or that equivalent findings or outcomes are obtained using either criteria. Decisions to be made include the selection of tests for best assignment composites and the determination of weights for these tests.

Procedures:

Project A's longitudinal data set includes specifically tailored CTP hands-on and job knowledge performance measures that were collected for 9 MOS, about 18 to 24 months after the cohort sample entered the Army. Operational ASVAB test and SQT scores for the same cohort sample were also obtained from official records. A 10th test, Assembling Objects, is also included as a predictor because it is being considered for use in ASVAB. The hands-on and job knowledge components of CTP are equally weighted. Only the skill knowledge component of CTP was administered to an additional 6 MOS. The data on the total 15 MOS enabled a comparison to be made between CTP hands-on and operational SQT job knowledge criteria.

Comparisons between the two criteria for seven types of indices are used to judge similarity of decisions or outcomes. These include: (1) similarity of tests and weights selected for assignment composites; (2) similarity in outcomes as measured by MPP, reliability of the criteria, stability of results using different data sets in cross samples; and (3) similarity of underlying dimensions as measured by intercorrelations between the two criteria, factor structure and loadings in the joint predictor-criterion space, and judgment of task overlap in the two criteria. The search for factor equivalence and construct validity was limited to predictor-dependent measures since the objective was to find the substitutability of either of the two criteria in developing and evaluating ASVAB classification procedures.

Findings:

The overall correlation between the two criterion measures is .46, while the correlation between the predicted performance scores of the two criteria is .92. These findings suggest that the two criteria measure different things, but when the criteria are measured in the joint predictor-criterion or valid space, either criterion could serve as a surrogate for the other in making classification decisions concerning ASVAB.

In considering the tests selected for the best 5-test composites, either criterion would select nearly the same tests for each job family test composite and provides comparable multiple validities.

The mean predicted performance (MPP) is .223 using CTP and .239 using SQT. Also, the stability of MPP in cross samples of different data sets is higher for SQT than for CTP. The pattern of test validities across ASVAB tests considered separately for the 15 MOS were also quite similar. The mean corrected validity for CTP is found to be .76, and for SQT the validity is found to be .83.

A principal component analysis of CTP and SQT in the joint predictor-criteria space yields six rotated oblique factors that showed good simple structure. Most of the 15 pairs of MOS had their highest coefficients (loadings) on the same factor for both criteria.

The overall conclusion, then, is that either criterion could serve as a surrogate for the other in conducting classification research on the ASVAB. Key decisions made using either a job knowledge or a job performance criterion were very similar and outcomes were judged to be equivalent.

ABBREVIATIONS

AO	Assembling Objects
ASVAB	Armed Services Vocational Aptitude Battery
AV	Assignment Variable
CE	Classification Efficiency
CTP	Core Technical Proficiency
EV	Evaluation Variable
LP	Linear Programming
LSE	Least Squares Estimate
MOS	Military Occupational Specialty
MPP	Mean Predicted Performance
NRC	National Research Council
PV	Predictive Validity
SME	Subject Matter Expert
SQT	Skill Qualification Test

THE SUBSTITUTABILITY OF CRITERIA IN THE DEVELOPMENT AND EVALUATION OF ASVAB CLASSIFICATION PROCEDURES

CONTENTS

Page

LIST OF TABLES

THE SUBSTITUTABILITY OF CRITERIA IN THE DEVELOPMENT AND EVALUATION OF ASVAB CLASSIFICATION PROCEDURES

Introduction

A. Need for the Study

In a related research effort the authors recommended new test composites for use in a proposed new two-tiered classification system for the Army (report in preparation). The characteristics of the proposed system are described below.

The use of an invisible or black-box first tier in which separate least squares estimates (LSE) assignment composites, based on the full set of Armed Services Vocational Aptitude Battery (ASVAB) tests, are computed for approximately 70 core MOS forming job families among the 250 entry-level MOS. A visible, revised second tier containing between 9 and 16 job families is provided for recruiting, counseling, use of minimum cut scores, and administration. The composite scores of the visible system are to be recorded on each soldier's individual record.

The remaining 180 or more jobs, not included in the study, would be linked to one of the 70 first tier families. Thus an expanded job family cluster is formed around each individual or core MOS identified as a single first tier family among the MOS for which data was available. Each such job family cluster is centered on a single kernel MOS or a core of MOS with adequate sample size to provide a stable LSE composite. Using the same linkage of the 180 jobs to the core jobs, the families of the second tier, or visible system, are formed.

The authors also suggest that the family clusters around each family core be refined as new classification information is systematically developed and also where data permit new core MOS be established, along with new family clusters. It was estimated that between 4 and 6 additional core jobs could be added to the current 66 core jobs to better represent the total MOS, and that a number of the existing 66 core jobs would be combined into the same family.

The proposed changes in the operational aptitude area composites were made based on the authors analysis of ARI's data set of ASVAB and SQT test scores limited to 66 MOS obtained during FY 1987-1989. However, before full consideration be given to replacing the current aptitude composites, the suitability of using SQT as a criteria measure of job proficiency needed to be carefully evaluated.

While hands-on performance measures are considered by many to be the only justifiable benchmarks of job proficiency, their costs in construction, administration, scoring and maintaining

physical conditions of measurement tools make them prohibitively expensive for use in classification research covering more than 250 entry-level jobs (MOS). Thus for classification research a surrogate is necessary. Paper-and-pencil job knowledge performance criteria would be considered adequate surrogates if it could be shown that the same decisions are reached or that equivalent findings or results are obtained for both types of criteria.

Fortunately, Project A's longitudinal data set includes both specifically tailored hands-on and job knowledge performance measures that were collected for 9 MOS during 1988 and 1989, about 18 to 24 months after the longitudinal cohort entered the Army. Operational SQT scores for the same cohort sample for the same years were also obtained from official records.

The equally weighted hands-on and skill knowledge components compose the core technical proficiency (CTP) criterion measure for 9 MOS while only the skill knowledge component of CTP for 6 additional MOS from Project A were utilized. This data set of 15 MOS, then, enabled a comparison to be made between the CTP criterion that we call in the present study the hands-on criterion and the operational SQT that we call the job knowledge criterion. The overall aim of the comparison is to determine if the two criteria are substitutable for one another in making decisions about ASVAB tests and one experimental cognitive test, Assembling Objects, in the classification process.

Details of the Project A's 7-year effort along with concurrent validity results for the 1983-1984 cohort sample were published in a special issue of *Personnel Psychology* (Project A, 1990). The intercorrelations between Project A's hands-on and job knowledge criteria along with an analysis of their validity patterns were reported in the National Research Council's (NRC) volume 1, *Performance Assessment for the Workplace* (Wigdor & Green, 1991). The NRC analysis directly relates to the suitability of using a skill knowledge criterion as a surrogate for a hands-on criterion in the selection context. Volume 1 consists of nine unsigned chapters by members of the committee on the Performance of Military Personnel, edited by Wigdor and Green. Volume 2 consists of eight chapters by individual authors, also edited by Wigdor and Green. For simplicity, Volume 1 is referred to here as Wigdor and Green and Volume 2 which indicates chapter authors is referenced using the author's name.

The present study examines the issue of the use of a surrogate criterion measure in the classification context using a longitudinal sample. The present analysis also uses a number of additional indices in comparing the two criteria in order to better understand the type and degree of equivalence and thus enhance our confidence in the use of a surrogate.

B. Some Observations on Hands-on and Job Knowledge Performance Measures

The "criterion problem" is often referred to as one of the most important and difficult problems in selection and classification research. Most investigators agree that even after seven decades of attention, it can still be characterized in the same way.

Wherry (1957), commenting on the lack of progress in criterion development compared to predictor development, called for the application of rigorous criteria for criterion developed. James (1973) and Smith (1976) suggested a construct validity approach for criterion development and also called for the same amount of effort be expanded for criterion measure development as had been expanded in the development of predictor tests. Gottfredson (1991) discusses a newer aspect of the criterion problem now referred to as criterion equivalence. She outlines a general approach for assessing the type and degree of similarity using a construct validity framework.

Work sample or hands-on measures are sometimes used as criteria in selection studies for training and for job proficiency. The most compelling argument for their use is that they have the highest level of fidelity for measuring criterion performance. As Borman (1991) points out: "What could be more direct and fair than to assess employees' performance on a job by having them actually perform some of the most important tasks associated with it?" (p. 305). Borman suggests that the evaluation of work samples as criteria is not quite so simple and involves consideration of such issues as test development, appropriateness of process and product measures and construct validity.

The major advantages of hands-on criteria is that they can provide concrete representation of jobs and thus have inherent credibility. The major disadvantages are that they can be very expensive to build, time consuming in testing, and that they can lose reliability in the field because of difficulties in maintaining exacting testing, scoring and physical conditions.

Job knowledge tests also are sometimes used as criteria in selection studies for training and for job proficiency. Job knowledge tests use some of the same content validation procedures that are used in the development of hands-on measures. Borman (1991) suggests that job knowledge tests may not be appropriate for tasks requiring fine motor coordination, quick reaction time and complex time-sharing psychomotor performance. Examples of such tasks are marksmanship, typing a letter under pressure and aircraft cockpit simulations, respectively.

Job knowledge tests are relatively inexpensive to construct, administer and score. Content validity can be demonstrated by an analysis of the degree of similarity between task elements in a job analysis and test characteristics. A strong relationship between job knowledge and job performance is sometimes expected on the basis of the assumption that "knowing" is necessary to performing or that knowing how and being able to perform are nearly the same thing.

Criticisms of job knowledge criterion measures include: contamination (the written format introduces unwanted verbal facility factors); deficiency (written tests do not directly measure the ability to perform); and method contamination (predictors and criterion may both be paper-and-pencil tests resulting in spuriously high correlations).

The NRC investigators state that hands-on performance measures can be considered benchmarks only to the extent they are valid and reliable measures of job performance constructs. Furthermore, threats to validity of criterion contamination and criterion deficiency apply as much to hands-on criterion measures as to alternate measures such as job knowledge (Wigdor & Green, 1991). However, Borman (1991) cautions that hands-on tests should not in any sense be considered as ultimate or even best criteria. He writes, "in keeping with Dunette's (1963) comments criticizing single, overall criteria measures, a single *method* of measuring performance should be discouraged" (p. 316).

Campbell et al. (1990) describe the goal of criterion development in Project A as the construction of multiple measures of critical components of job performance to cover the total performance domain for a representative sample of entry-level enlisted jobs. Results revealed a five-construct, multidimensional representation of job performance. Only the job-specific core technical proficiency (CTP) dimension, from among the five criterion constructs, differed significantly across jobs to warrant their use in a classification process. The NRC investigators appear to agree, at least to some extent, with Dunnette's quote when they write that since there is no empirical standard or "ultimate" criterion against which to validate criterion measures, job performance can be measured in many ways.

C. Research Objectives

The overall objective of the present research is to determine the adequacy of operational SQT measures of job knowledge to serve as surrogates of CTP measures of hands-on job proficiency in developing classification procedures using ASVAB and the AO test as assignment variables (AV). The assignment variables are based on least squares estimates (LSE) of SQT or CTP scores, using either the

full set of tests or the best five-test composites. The assignment variables are predicted performance measures that are used in optimally assigning individuals to job families.

The SQT criterion would be considered adequate if it could be shown that the same decisions are reached or that equivalent findings (outcomes) are obtained for both SQT and CTP. More specifically, the decisions to be made include: the selection of tests for the best weighted 5-test composites, and the determination of tests weights in the 5-test composites or in the full set of tests.

The comparability of results or outcomes is measured by mean predicted performance (MPP) after simulated assignments to job families are made. The stability of MPP results are further evaluated in cross samples using different data sets. Reliabilities are also compared. The type and kind of similarity of the two criteria are judged by the intercorrelations of the two criteria and by factor structure and loadings of the criteria in the joint predictor-criterion space, as defined by predictor-dependent measures.

The objectives of this research are limited to determining if SQT is an acceptable surrogate for CTP in making decisions on classification procedures in the context of using ASVAB and AO predictors. While the SQT measure has been accepted in practice as relevant to some personnel goals of the Army such as its use in retention and promotion of individuals, we wish to determine the extent to which it also can be validly and reliably employed as a surrogate for the more widely accepted hands-on measure. The broader goal of fully examining the degree of criterion equivalence of the two measures based on the scientific principles of construct validity as described by Cronbach and Meehl (1955) requires a much different research strategy.

D. Relevant Military Studies

One approach that can be used in evaluating alternative criterion measures stresses a specific application and raises the question as to whether one criterion measure can serve as a surrogate for another measure. Here one may ask what differences in decisions are made in validating selection and classification procedures for ASVAB. If it is agreed, for example, that hands-on measures are the benchmarks for measuring job proficiency, can less costly job knowledge measures be substituted to make comparable decisions in selecting and weighting test composites or in obtaining comparable outcomes in evaluating predictor composites?

The focus, then, of this preceding approach is not on an assessment of the type and degree of similarity of factor structure and construct validity but on decisions reached using one or the other of the criterion measures an on outcomes. This is the approach used by the NRC investigators in analyzing the quality of hands-on and job knowledge measures in ASVAB criterion-related studies (Wigdor & Green, 1991) pp. 47-152. As Gottfredson (1991) concludes in discussing the bottom line of nonequivalences among alternative criteria, "Two measures are substitutable for a given purpose when their estimated utilities are the same and when these estimates are made with equal confidence, even though many particular facets of those measures may differ" (p. 87).

A second approach in evaluating alternative criterion measures stresses an in depth study of factor structure, construct validity and relevance (value to the organization) of the underlying dimensions being measured. Gottfredson (1991) provides a cogent analysis of the major issues that should be considered in determining the equivalence of criteria of the same general type used for the same purpose. Wherry, Ross and Wolins (1956), use a factor analytic approach in judging similarities using military job performance data.

We begin our review of these two approaches by referring to the criterion related data provided in the National Research Council's on *Performance Assessment for the Workplace,* Volume 1, edited by Wigdor and Green (1991). The overall point of view clearly stressed in this volume is to bring to the workplace the "authentic assessment" approach advocated by the movement to reform education. Wigdor and Green write:

> This approach eschews written multiple-choice tests in favor of demonstrated performance on tasks in which desired knowledge and skills are used. The emphasis is on giving pupils the opportunity to demonstrate what they can do, rather than how well they can answer questions about a subject. (p. 1)

A problem of particular concern for the NRC investigators in considering paper-and-pencil job knowledge tests as possible surrogates for hands-on tests occurs when a written multiple-choice criterion measure is used for validating ASVAB, also composed of paper-and-pencil multiple-choice predictor tests. This "method contamination," as many would agree, could result in spuriously high validities, since both the predictor and the criterion measure are similarly contaminated by a verbal ability facility, an ability that may not be a part of the job or even an object of measurement (p. 150).

Keeping this precaution in mind, the investigators compared the uncorrected predictive validities of the operational aptitude area composite on the ASVAB using job knowledge tests and hands-on performance measures for 9 Army and 4 Marine Corps entry-level jobs. The Army data were obtained from Project A's concurrent validation in 1983-1984 and the Marine Corps data were based on tables provided in 1988.

The NRC investigators found that the validities were higher for the job knowledge criterion measures than for the hands-on criterion measures in all 13 jobs, a finding consistent with expectations based on the argument of common method variance. The differences between the two criterion measures for the Marine Corps jobs averaged about .06 with the largest difference being around .08. For the Army jobs, however, some differences were relatively larger. The investigators concluded that the operational composites had useful validities for both criterion measures in all 13 jobs and that validation studies using job knowledge as criterion measures can provide useful indicators of the predictive validity of ASVAB scores for military jobs.

It is noteworthy that in the NRC analysis, operational aptitude area composites were employed to make the comparisons between the two criteria. Such composites generally yield much lower validities than LSE composites of the criterion, based on the full set of ASVAB predictors as found in Project A by McHenry et al. (1990). However, the NRC analysis, as the investigators state, does not indicate whether the use of one or the other of the two criteria would lead to different decisions about the best individual ASVAB predictors to use in selecting for a specific job. The present study will address the magnitude of predictions, the similarity of decisions reached and the outcomes of decisions employing both individual predictors and composites within the context of classification.

McHenry et al. (1990) estimated that "method variance" accounted for .16 correlational points of the mean shrunken *R* of .62 between the written skill knowledge tests and ASVAB predictors. The investigators obtained this result by removing method factor variance from the core technical proficiency factor variance. McHenry et al. suggest that the term "method factor" itself may be a misnomer. Since "the written test factor may reflect comprehension of the manuals, instructions, and other materials that must be read on the job. For several of the jobs that were studied, excerpts from technical manuals and other learning aids were incorporated by design into the written job knowledge tests" (p. 348). The findings in the McHenry et al. study indicate that even if method variance is considered undesirable, the desirable, valid, variance remains quite substantial, retaining about 93 percent of the total variance.

Paper-and-pencil job knowledge criteria may be both deficient and contaminated. The most apparent deficiency is that they do not directly measure actual performance on a task. As noted earlier, the most apparent contamination is that both the criterion and the ASVAB predictors use a similar written word format. Correlations between job knowledge tests and hands-on tests indicate the degree to which they measure the same constructs or dimensions. The relationship between the two types of measures across different jobs, then, answers to some degree concerns about both deficiency and contamination. The NRC investigators reported correlations between job knowledge and hands-on measures for 15 entry-level jobs (the same 9 Army and 4 Marine Corps jobs that were used to obtain the validity results reported in the preceding section, plus 2 Navy jobs). Uncorrected correlations were found to range from .35 to .61; the investigators note that if both criteria were corrected for attenuation, the overall degree of relationship would be even stronger (Wigdor & Green, 1991).

Since the correlations found between job knowledge and hands-on job performance tests are substantially less than 1.0, the NRC investigators concluded that the two criteria do not measure the same dimensions. They note:

> In other words, using a strict standard of equivalence, job knowledge tests are not interchangeable with hands-on performance tests. Compared with other variables, however, the link between the two types of measures is relatively strong. If it could also be shown that decisions about the choice of predictor variables and the rules used for selection and classification would be unchanged due to the choice between these two types of criterion measures, then a case might be made that paper-and-pencil job knowledge tests are adequate surrogates for the more expensive hands-on performance tests. (pp. 151-152)

Hunter (1983, 1986) reported corrected correlations between job knowledge and hands-on as high as .80 and suggested that very high correlations between the two performance are to be expected. Hunter argued that knowing how and being able to do something are about the same. Wigdor and Green (1986), however, point out that written tests require more of an inferential leap from test to job performance than do job-sample tests.

In our analysis, in addition to examining the direct empirical relationship between the two types of criteria, we explore the relationships between the two criteria in the joint predictor-criterion space or, in other words, determine the correlations of predicted performance scores. The correlations of predicted performance scores indicate the degree of similarity between the two criteria when considering only the valued variance portion of each score.

The present study also shows the predictor tests selected for use in assignment composites in classification procedures rather than showing the predictor test selected for use in selection procedures as is conventionally reported..

We next turn in our review to the approach that focuses on factor structure and construct validity in evaluating alternative criterion measures. As mentioned earlier, Gottfredson (1991) provides a review of issues involved in the process of assessing equivalences and develops an outline of a strategy for obtaining evidence in judging similarities. Her article includes a detailed analysis of Wherry et al.'s (1956) study on "similarity coefficients" of correlational data because Gottfredson felt it was useful in revealing correlational methods in determining degree of equivalences and in pointing out potential limitations of these methods. According to Gottfredson, Wherry et al. examined variations in seven indices:

(1) the magnitude of the criterion intercorrelations corrected for attenuation;
(2) the similarity of the profiles of factor loadings based on a joint analysis of criteria and predictors;
(3) the similarity of the profiles of factor loadings based on an analysis of predictors only, with the criteria added by extension;
(4) the overlap of elements checked as present in the criteria on some list of job elements;
(5) the similarity of the profiles of criterion-predictor correlation coefficients;
(6) the similarity of the profiles of criterion-predictor beta weights (standard score regression weights); and
(7) the relative success of cross-validation and criterion extension for a pair of criteria (the success of betas from another criterion compared with that for betas from the criterion itself, where both sets of betas come from a previous sample).

One serious limitation of several factor analytic indices employed in the Wherry et al. study, according to Gottfredson, is that they are entirely predictor-dependant and do not include data about the direct relations of the criterion measures with each other. She emphasizes that predictor-dependent indices can only reflect commonalities in the criteria that are present in the predictor. A second limitation Gottfredson points out concerning comparisons of criterion measures based on factor analytic approaches is that equivalences will change depending on the set of predictors and criteria data used in the analysis. She also notes that neither beta weights nor factor loading indices are appropriate for determining degree of factorial equivalence.

In the present analysis we employ several predictor-dependent indices for the limited objective of finding a surrogate criterion measure in determining and evaluating ASVAB classification procedures. We believe that factor analytic approaches in the joint predictor-criterion space used in our study in

conjunction with judging similarities in intercorrelations between the two criteria, and in predictive validity patterns, as well as an examination of tests chosen for assignment composites, contributes to both our understanding of type and degree of similarity and to our confidence in the use of either of the two criteria as a surrogate for the other.

Procedures

A. General Approach

This study consists of a series of comparisons between two criteria, core technical proficiency (CTP) and skill qualification test (SQT) performance measures, utilizing a number of indices as a basis of determining criterion equivalence. The indices used in judging the similarity of the two criteria are: criterion intercorrelations; tests selected for assignment composites, and the beta weights of tests selected for the composites; mean predicted performance (MPP) results; stability of results in cross samples using different data sets; factor structure and loadings; reliability of the criteria; and overlap in task elements in the criteria.

Classification efficiency is measured in terms of mean predicted performance (MPP) after simulated optimal assignment of recruits to job families are made. Optimal assignments are made using weights for the assignment variables (AV) obtained in an analysis sample. Assignment variables, based on the least squares estimates (LSE) of CTP or SQT scores using the full set of ASVAB tests and the Assembling Objects (AO) test, or on the LSE of the best five-test composites, constitute the predicted performance measures. The best five-test composite for each job family is the composite having the highest *R* for that job family from among all possible 5-test combinations.

The present study employs the data set of Project A: Building the Career Force. In the second phase of Project A, the ASVAB tests and a set of experimental predictors were administered to a longitudinal sample of recruits at the time of their entry into the Army during 1986-1987. The hands-on component of the CTP criterion was administered to this sample for 9 MOS during late 1988 and early 1989, about 18 to 24 months after entry. Only the job knowledge component of the CTP criterion was used for an additional 6 MOS. The skill knowledge written component of CTP was administered earlier, during end-of-training testing. SQT scores were obtained from the official records of the same cohort sample. For the longitudinal analysis, 21 jobs out of more than 250 entry-level jobs (MOS) were chosen for study on the basis of sample size and criticality to the Army. Of these 21 jobs, we selected for the

present study 15 jobs, including all 9 jobs using hands-on measures, on the basis of sample size considered adequate for classification research.

B. Core Technical Proficiency Data

The criterion development effort of Project A focused on four major measurement methods: hands-on job sample tests, multiple-choice job knowledge tests, peer and supervisor ratings, and administrative data from existing files (C. H. Campbell et al., 1990). The CTP criterion measure was composed of a "hands-on" component and a paper-and-pencil job knowledge component. Because of the time and expense involved, hands-on components were developed for only 9 jobs. The CTP measure for each of these 9 jobs also included a paper-and-pencil job knowledge component. This group of jobs was designated by Project A researchers as Batch A. Only the job knowledge component of CTP was available for the other six jobs used in the present analysis. This group of jobs was designated as Batch Z.

The CTP criterion can be described as a measure of the proficiency with which the soldier performs the task that are critical to the MOS. Such tasks represent the core of the job and they are the primary definers of the MOS. CTP, then, measures how well the soldiers can execute the core technical tasks the job requires, given a willingness to do so (Campbell and Zook, 1994).

The construction of the CTP measures involved a rigorous sequence of steps based, in part, on job analysis information and manuals describing specific tasks that incumbents in MOS must perform. Subject matter experts (SME) developed descriptions of all major tasks that compose each job. SMEs also provided judgments on task clustering, importance and difficulty. A task selection panel designated about 15 tasks on each MOS that should be included in a job specific proficiency measure.

Each task consisted of a number of scorable steps so that the set of tasks for an MOS could be completed by an examinee in four hours. Eight NCOs were used as scorers. Both the hands-on and job knowledge components used some of the same task analysis information. A multiple-choice format was used for the job knowledge tests.

In the present study, measurement of job proficiency for the 9 Batch A MOS was composed of two unit-weighted components: MOS-specific hands-on and MOS-specific job knowledge measures. Measurement of job proficiency for the 6 Batch Z MOS was composed of only one component: the MOS-specific job knowledge measures. A detailed description is provided by C. H. Campbell (1990)

and a brief outline of the steps in development of the hands-on tests is provided by Wigdor and Green (1991), pp. 69-71.

C. Skill Qualification Test Data

As noted earlier, there are about 250 entry-level MOS in the Army. Each of these MOS is composed of one to five skill levels, with skill level 1 being the lowest and including paygrades E-1 through E-4. Prior to 1983, the SQT had both written and hands-on components measuring job proficiency in an MOS. After 1983 the SQT was designed only as a task-based paper-and-pencil test of job proficiency. The SQT program was cancelled in 1991. Soldiers were required to take the SQT annually after completing 11 months or more of service. In the present study, SQT scores for FY 1988-1989 were obtained from official records of the cohort sample for the same time frame used to collect the CTP data. These SQT years were considered by ARI to be psychometrically good SQT years in terms of discriminability and reliability of the measures.

Like the job knowledge measures of CTP, the SQT for an MOS were composed of significant task elements. The tasks to be measured were selected by SME using job analysis information and manuals. Test questions were standard four alternative multiple-choice items, with one correct answer. The tests were administered through a local Training Standards Office once each fiscal year during a three month Army-wide SQT period. The SQT scores were utilized in considering soldiers for promotion to the next higher skill level and to identify those that could be barred from reenlistment. A report by R. C. Campbell (1994) provides a history of the development and use of SQT.

D. Predictor Data

The predictors consisted of the 9 ASVAB tests and the Assembling Objects (AO) test, a spatial ability measure, that may be added to the ASVAB.

The ASVAB tests and AO were standardized to have a mean of zero and a standard deviation of one in the Army input sample. The set of CTP and SQT scores in each of the 15 MOS were also standardized to have a mean of zero and a standard deviation of one, but calculated within a single MOS.

The validities in this study were corrected for multivariate restriction in range separately by MOS. Range restriction was due to operational assignment effects, the restriction in range impact of assignment to MOS from a common entry pool. Since this study uses the Army input sample rather than

the youth population as the basis for making this correction, no further correction is made for restriction due to selection effects. Validities are also corrected for unreliability of the criterion variable prior to the restriction in range correction.

E. Simulation of the Operational Classification System

The present study is accomplished within the classification context and this section highlights some concepts of classification and of system simulation. A detailed discussion is given in Johnson and Zeidner (1991). Potential classification efficiency is estimated by the simulation of a system in which the assignment of a recruit to a job family is made in such a way as to optimize the sum of all recruits' AV corresponding to the family to which each person is assigned. A linear programming (LP) algorithm is used to maximize this total sum of AV as the objective function. This maximization of the objective function is accomplished under the constraint of meeting quotas for each assignment target set proportionately to the operational accession numbers for the 66 MOS included in the analyses. Optimal assignments to maximize the overall objective function can also be constrained to maintain a minimally acceptable objective function value in each job family. This constraint was not applied in the present study.

Assignments to job families are determined in one or more simulation subsamples used as experimental cross samples. In the most highly biased design to estimate MPP, the same data set may be used for the analysis, evaluation, and simulation subsamples. However, in a completely unbiased design, a triple cross-validation design is utilized. The triple cross-validation design calls for the use of two independent subsamples for the computation of regression weights for AV and for EV. A third independent subsample is used as the source of the score vectors (entities) to which these weights are applied to obtain new and totally unbiased AV and EV. This third sample is used for the actual conduct of the simulation and evaluation processes. The analysis sample data used to compute regression weights for either AV or EV are corrected first for attenuation and then for a restriction in range effect caused by the classification and assignment process. The test score vectors constituting entities in the cross samples used for simulation are not corrected in any way.

In the research design, which uses empirical test scores instead of synthetic test scores, we do not correct intercorrelation and validity matrices obtained in the analyses samples for restriction in range selection effects, but, as noted earlier, do correct for restriction in range due to assignment effects using a common entry pool. Also as noted, validities are corrected for unreliability of the criterion variables.

The regression weights applied to entities in the simulation sample to obtain both AV and EV are computed from these corrected matrices.

Classification efficiency (CE) is measured by mean predicted performance (MPP) after optimal assignment. CE is conducted using predicted performance (i.e., the evaluation variable) based on the same set of prediction variables used to compute AV. A proof of the adequacy of substituting predicted performance for actual performance in evaluating CE was shown by Brogden (1955).

The generic simulation paradigm can be described in terms of the following roles given to empirical samples of entities: (1) the analysis role in which MOS are clustered into job families, weights for AV are computed and tests selected for use in composites; (2) the evaluation role in which weights are computed for LSE, using all predictor information and SQT scores in each MOS or job family to provide the best estimate of criterion scores; and (3) the simulation role in which entities are optimally assigned to job families (using weights for AV computed in an analysis sample and MPP computed using the weights for evaluation variables (EV) computed in the evaluation sample.

Predictor intercorrelations and validities are computed within MOS samples and then aggregated to provide similar information on job families. Only the simulation sample utilizes individual entities. Entities consist of the predictor score vectors (without criterion scores or MOS identification). AV and EV are computed for each entity in order to assign entities to job families and to evaluate results of the assignments in terms of MPP.

F. Research Design

In the present study the AV and EV weights were based on the same sample and the cross samples also partially overlapped the other two samples. Figure 1 shows the research design employed in the present study. It indicates that the validities determined in the EV data set are used to compute the job family weights determined in the AV data set. Note that EV and AV samples employ the same subjects (N = 6,809 when only the 9 ASVAB tests are used or N = 5,364 when the AO is also utilized). Because there is not a complete separation between the job weights and family weights in the two identical samples, the results are partially biased. Also, since the evaluation weights are also applied to many of the same entities or subjects in the cross samples, with a 73 percent overlap between the cross samples and the evaluation samples, the results again would be partially biased. However, in the case of the cross samples, the entities were randomized before being assigned to a cross sample and optimally

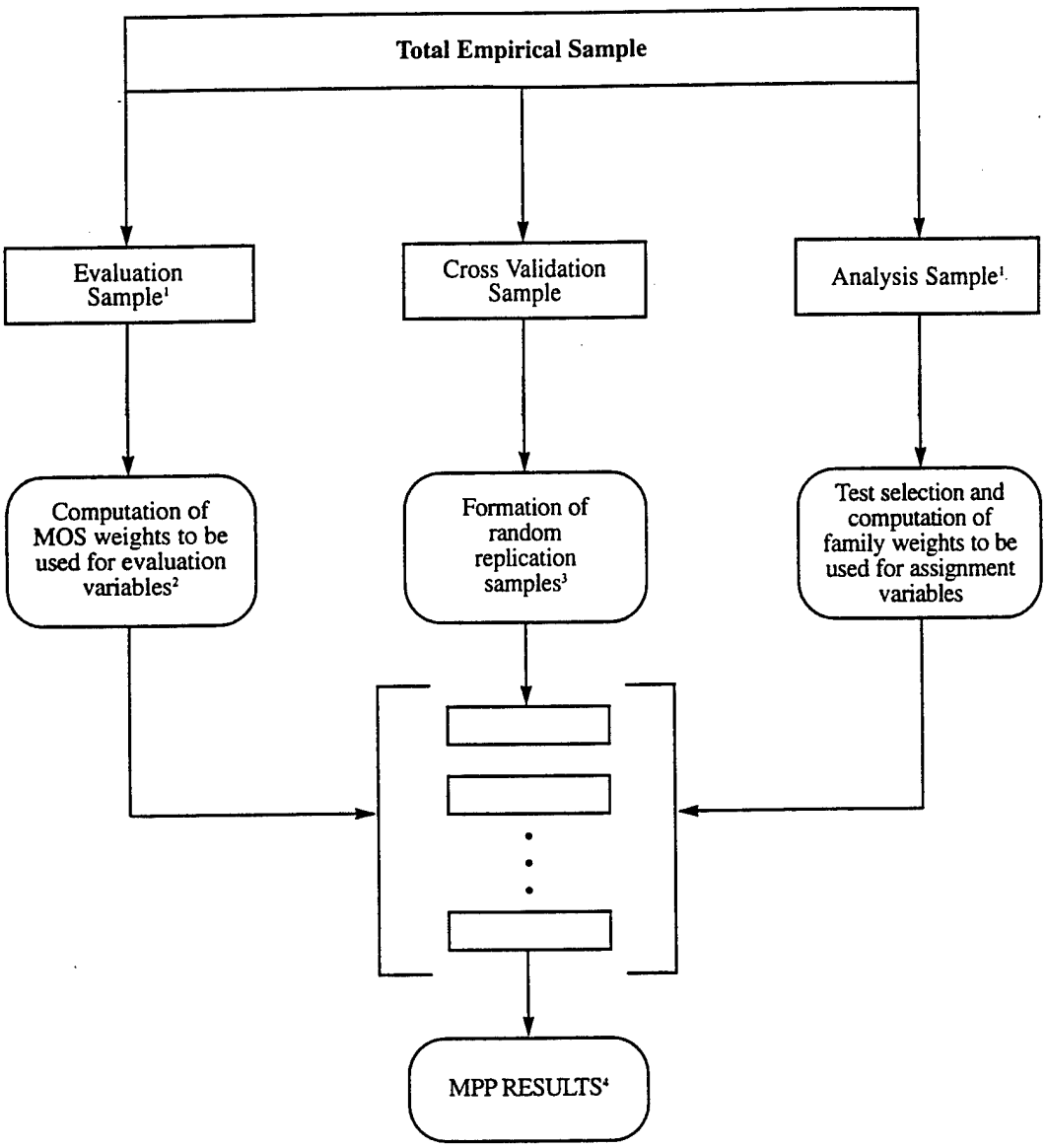

Figure 1. Research design. [1]Evaluation sample and analysis sample sizes were equal. [2]Evaluation job weights were computed from Project A empirical sample. [3]Sample size of assigned individuals numbered 238 in each cross sample; in the aggregate, N numbers 5,364 for each condition. [4]Mean predicted performance was computed after simulation of the assignment process using the same evaluation variable and same weights for each job across all experimental conditions.

assigned to a family. An analysis showed that about 10 percent of the entities in the cross samples were found to be assigned to the same MOS as in the EV sample, a number that could be expected by chance. It is important to note that there is no reason to expect a differential effect on results for either of the two criteria because of the use of non-independent samples. The focus of this study, the type and degree of similarity of the two criteria, remains unbiased; only the magnitude of MPP is raised proportionately for both criteria.

Results

A. Predictors and MOS

Table 1 shows the current 9 ASVAB tests; the tenth predictor test included in this study as noted earlier is the Assembling Objects (AO), which include 36 items and takes 18 minutes. Table 2 gives the composition of each unit-weighted operational aptitude area composite. We include a description of these composites as a point of comparison with the test utilized in the five-test assignment composites (AV). As noted earlier, two types of AV are employed, the LSE of the complete 10 tests for each job family and the best 5-test LSE composites for each job family. Again, "best" refers to the 5-test composite having the highest multiple correlation (R) for each job family from among all possible 5-test combinations. While the operational job family structure is retained in the present analysis, new LSE test composites are used to assign individuals to job families. Of the 15 MOS included in the present study, none represents the Electronics (EL) job family. This reduces the operational job families from 9 to 8 families in the present analysis. Again, no use is made of the operational test composites, but the operational family structure is retained in this study.

In Table 3, we show the MOS included in the present study by category type (Batch A or Batch Z) and sample sizes. These are the MOS we used from the longitudinal data set and they vary slightly from the sample of jobs. Also there are slight changes in the way CTP is measured here compared to the way it was measured in the earlier and better CTP measures used in the earlier and better known concurrent validation data set.

Table 1

ASVAB Tests (Forms 8-22)

Subtest	Number of Items	Time in Minutes
1. General Science (GS)	25	11
2. Arithmetic Reasoning (AR)	30	36
3. Verbal Ability (VE)*	50	24
4. Numerical Operations (NO)	50	3
5. Coding Speed (CS)	84	7
6. Auto & Shop Information (AS)	25	24
7. Mathematics Knowledge (MK)	25	11
8. Mechanical Comprehension (MC)	25	19
9. Electronics Information (EI)	20	9
Total	289	144

* A composite of Word Knowledge & Paragraph Comprehension tests

Table 2

Current Army Operational Aptitude Composites of the ASVAB

Code	Composite Name	Definition
CL	Clerical Administration	VE + NO + CS
CO	Combat	AR + CS + AS + MC
EL	Electronics	AR + MK + EI + GS
FA	Field Artillery	AR + CS + MK + EI
GM	General Maintenance	MK + EI + AS + GS
MM	Mechanical Maintenance	NO + AS + MC + EI
OF	Operators & Food	NO + AS + NC + VE
SC	Surveillance & Communication	AR + AS + MC + VE
ST	Skilled Technical	GS + VE + MK + MC

Table 3

Project A Military Occupational Specialties (MOS) and Sample Sizes Using the Longitudinal Validation Data Set

Job (MOS)	Batch A MOS	*n*	Job (MOS)	Batch Z MOS	*n*
11B	Infantryman	538	12B	Combat Engineer	465
13B	Cannon Crewmember	579	16S	Man Portable Air Defense Crewmember	264
19K	COM-1 Abrams Armor Crewmember	366	54B	NBC Specialist	286
31C	Single Channel Radio Operator	140	55B	Ammunition Specialist	123
63B	Light Wheel Vehicle Mechanic	386	76Y	Unit Supply Specialist	385
71L	Administrative Specialist	284	94B	Food Service Specialist	462
88M	Motor Transport Operator	240			
91A	Medical Specialist	564			
95B	Military Police	282			

Note. In Project A, Batch A represents jobs with hands-on, job-specific criterion measures and Batch Z represents remaining jobs for which no hands-on, job-specific criterion measures were obtained. Enlistees in both batches were administered paper-and-pencil tests of job knowledge and training achievement tests. MOS = Military Occupational Specialty.

B. Relationship Between CTP and SQT

Table 4 presents the mean correlations between CTP and SQT criteria across jobs. First looking at the 15-job conditions, we find that the correlation between the two criteria for total test scores to be .487 after correction for criterion attenuation and multivariate restriction in range for the Army input population. The correlation of .487 demonstrates that while job knowledge measures are correlated with hands-on performance measures, the two criteria do not measure the same dimensions.

Table 4

Mean Correlation Between CTP and SQT By Job for Total Test and Predicted Performance Scores

Space	15 Jobs	9 Jobs
Total Test	.487	.460
Joint Predictor- Criterion	.926	.916

Note. Averages were computed across fifteen jobs used in Project A and across ten jobs for the 10-test battery. CTP = core technical proficiency; SQT = skill qualification test.

However, we find the correlation for predicted performance scores to be .926 after correction for attenuation and range restriction. In using predicted performance measures, we are examining the relationship between the two variables in the predictor-criterion space or the relationship of the two variables considering only the valid components of total test scores. The mean correlations between the predicted performance scores of CTP and SQT provides an indication of overlap in the predictor-criterion space (valid space), as defined entirely by the relationships between the 10 predictor tests and each of the two criteria. The correlation of .926 between predicted performance scores clearly indicates that the two criteria are interchangeable for practical purposes in terms of the validities of ASVAB tests and the AO test. This finding alone begins to suggest that with regard to making decisions about ASVAB tests such as consideration of adding a new cognitive test to the Battery and the number and composition of tests for use in assignment composites and in specifying test weights, the SQT measure appears to be a promising surrogate for the hands-on CTP measure.

Table 4 also presents the mean correlations across jobs for the 9-job condition (Batch A). As noted, Batch A incorporates into the CTP criterion equally weighted hands-on and skill knowledge components. The 6 jobs (Batch Z) measuring only the job knowledge component in the CTP, are not included in the 9-MOS analysis. Thus the analysis of the 9-MOS condition provides a more rigorous separation between the hands-on and job knowledge criteria. The overall correlation for the 9-MOS condition, then, shows the relationship between the full CTP measure (hands-on and written job knowledge components) with the operational SQT job knowledge measure. The total test scores of the two variables to be .460 and between the predicted performance scores is .916.

In comparing the 15-MOS and the 9-MOS conditions we find very similar results. For example, the mean correlation in the joint predictor-criterion space is .926 for the 15-job condition and .916 for the 9-job condition. Significantly, the addition of the hands-on component to the job knowledge component adds only .01 correlational point, although the effects of number and type of jobs and difference in criteria can not be separated in this analysis. Results for the total test score are similarly, although there is a difference of about .03 correlational points between the two conditions.

C. Test Selection and Weights

Table 5 presents a comparison of CTP and SQT criteria on 8 job family composite validities and test beta weights. As noted earlier, the EL family was not included in this data set. The multiple correlation for each job family is computed using the best-weighted full ten-test composite. The mean multiple *R* for CTP is .578 and for SQT is .552. Several of the job families have large differences in multiple validities for the two criteria. For example, CO has a .20 point difference in favor of CTP and

MM has a .14 point difference in favor of SQT. It should be noted that these are back multiple correlations capitalizing on sampling error; cross sample correlations would shrink and the differences in correlations between the two criteria may be smaller, but the obtained differences are trivial. The multiple correlations found here between the predictors and each of the criterion measures provide a considerable degree of similarity in results or outcomes using either criterion. This type of analysis is analogous to most ASVAB selection studies that implicitly assume independent pools of applicants. Predictive validity, of course, is not the measure of classification efficiency we employ in this study.

Table 5

Comparison of CTP and SQT for Job Family Composite Validities and Test Beta Weights for 10-Test Composites

			Test									
Job Family	Criterion	Predictive Validity	GS	AR	NO	CS	AS	MK	MC	EI	VE	AO
CL	CTP	.593	.150	.164	-.015	.100	-.004	.179	-.052	.028	.123	.138
	SQT	.515	.136	.191	.027	.078	.000	.203	-.057	.011	.035	.078
CO	CTP	.674	.061	.034	.027	-.037	.158	.026	.119	.099	.069	.299
	SQT	.474	.026	.095	.039	-.041	.097	.117	-.003	.106	.021	.170
FA	CTP	.419	.049	.060	.033	.010	.045	.060	.117	.020	-.004	.177
	SQT	.426	.147	.129	.006	-.072	-.008	.054	.009	.093	.091	.017
GM	CTP	.660	.479	-.002	.086	.086	.182	-.009	-.038	-.035	-.071	.256
	SQT	.538	.274	.037	-.049	.229	.108	.094	-.157	-.126	.088	.245
MM	CTP	.609	.050	.095	.109	-.001	.361	.009	.023	-.057	.077	.212
	SQT	.744	.101	.155	-.026	-.069	.379	.091	.133	.068	-.081	.093
OF	CTP	.566	.006	.115	-.016	-.013	.066	.060	-.004	.064	.190	.243
	SQT	.539	.073	.089	.017	.008	.094	.040	.007	.038	.122	.241
SC	CTP	.508	.012	.065	.104	-.242	.399	.149	-.054	-.109	.085	.038
	SQT	.593	-.022	.156	.016	-.156	.147	.229	.126	.166	.162	.066
ST	CTP	.597	.061	.049	-.035	.098	.046	.038	.035	.087	.162	.274
	SQT	.585	.047	.142	-.007	.068	.152	.082	.070	.080	.040	.152

Note. CL = Clerical/Administrative; CO = Combat; EL = Electronics Repair; FA = Field Artillery; GM = General Maintenance; MM = Mechanical Maintenance; OF = Operators/Food; SC = Surveillance/Communications; ST = Skilled Technical; GS = General Science; AR = Arithmetic Reasoning; WK = Word Knowledge; NO = Numerical Operations; CS = Coding Speed; AS = Auto Shop Information; MK = Mathematical Knowledge; MC = Mechanical Comprehension; EI = Electronics Information.

In examining the 10 pairs of beta weights for each of the 8 job families in Table 5, we find 18 of 80 comparisons have beta weights larger than .10 and of these only a few are found to be larger than .15. Although it can be seen that the betas in the table "bounce" a bit as they customarily do, they still provide comparable multiple correlation values in the comparisons of CTP and SQT. For each job family, the two sets have either the highest or next highest beta weights for the same two tests for each of the two criteria. The data also indicate that the experimental AO test appears to be a good measure of *g* with

relatively high weights on five of the eight test composites. (For classification purposes, a test with greater differential validity would be preferable. However, adding an additional test is useful for increased classification efficiency.)

In Table 6 the multiple *R*, the selected tests for each composite, and weights for the best 5-test composites are given. Best, as noted earlier, refers to the 5-test composites resulting in the highest multiple *R* for each job family from among all possible 5-test combinations. In a previous study, Scholarios, Johnson and Zeidner (1994) found predictive validity (PV) to be as good as Horst's index of differential validity (H_d) for selecting tests for a composite. However, H_d was found to be better for selecting tests for a battery.

Table 6

Comparison of CTP and SQT for Job Family Composite Validities and Test Beta Weights for the Best 5-test Composites

			Test									
Job Family	Criterion	Predictive Validity	GS	AR	NO	CS	AS	MK	MC	EI	VE	AO
CL	CTP	.586	.229	.165		.101		.181				.128
	SQT	.513	.140	.187		.096		.211				.057
CO	CTP	.671					.154		.143	.117	.108	.319
	SQT	.472		.099			.102	.134		.118		.168
FA	CTP	.416		.067			.056	.084	.127			.183
	SQT	.421	.162	.129				.049		.098	.087	
GM	CTP	.657	.404		.090	.081	.155					.227
	SQT	.518	.225			.214		.066			.068	.201
MM	CTP	.608		.104	.108		.352				.095	.221
	SQT	.735	.101	.182			.396		.160			.085
OF	CTP	.564		.142			.060			.072	.204	.249
	SQT	.537	.091	.118			.096				.130	.256
SC	CTP	.446	.031	.124	.016		.339				.049	
	SQT	.571		.137			.139	.208		.142	.135	
ST	CTP	.593		.074		.070				.139	.211	.307
	SQT	.577		.159			.171	.112		.126		.192

Note. See Table 5 for abbreviations.

The mean multiple correlations are .568 for CTP and .543 for SQT. We find that 26 of 40 pairs of comparisons of tests selected for inclusion in the best 5-test composites were the identical (matching) tests for each of the two criteria. The mean beta weight for the 28 non-matching tests for the two criteria is .106 compared to the mean beta weight for the 52 matching tests of .197. Considering that beta weights can be very unstable in some conditions due to sample size and number of correlated predictors,

we judge that the betas are surprisingly stable for the criteria collected at different times and under such different conditions.

Table 7 provides another indication of the general similarity of test selection for job family composites with the use of either CTP or SQT. The table shows the number of times each test is chosen as one of the best tests in a 3- or a 5-test composite. For example, AR is selected as one of the five best tests for 6 of the 8 job families using CTP as the criterion and 7 of the 8 job families using SQT. The NO test, however, is selected for 3 of the 5-test composites using CTP, but with quite low beta weights, and for none of the composites using SQT. The MK test is selected for only 2 of the 5-test composites using CTP and for 6 of the composites using SQT; in the case of MK, however, the beta weights are relatively high using SQT. These are instances of dissimilarities of decisions, but the overall outcomes are comparable.

Both the results for the 10-test and 5-test composites support the substitutability of one criterion measure for the other in terms of similarity of outcomes. The results for the 5-test composites provide additional strong evidence of similarity of decision in choosing tests for assignment composites.

Table 7

Number of Times Test Appeared in 5-test Family Composites

Test	criterion	Times in top 1-3 tests	Times in top 1-5 tests
GS	CTP	2	3
	SQT	3	5
AR	CTP	3	6
	SQT	5	7
NO	CTP	1	3
	SQT	0	0
CS	CTP	0	3
	SQT	1	2
AS	CTP	4	6
	SQT	3	5
MK	CTP	2	2
	SQT	3	6
MC	CTP	2	2
	SQT	1	1
EI	CTP	1	3
	SQT	3	4
VE	CTP	3	5
	SQT	1	4
AO	CTP	6	7
	SQT	4	6

Note. See Table 5 for abbreviations.

D. Classification Efficiency

Table 8 provides results in the classification context, the principal focus of the present study. Classification efficiency is measured by MPP in a simulation process described in the Procedures section. The table shows MPP for 8 assignment conditions. First, examining the results for the full 10-test composites across the 15 MOS, the MPP after optimal assignment is .237 using CTP and.242 using SQT. Examining the results for the 5-test composites across the 15 MOS, the MPP is .221 using CTP and .220 using SQT. Thus we see that the findings for both the 10- and the 5-test composites using either criterion are very comparable, the differences between the two criteria being .005 and .001 of a standard deviation unit, respectively. However, while the overall mean difference using either criterion were quite small, a number of job families had large disparities in MPP. At this point of study we have no explanation for these disparities.

Table 8

Mean Predicted Performance for 8 Assignment Conditions: Classification Effect Only

	15-MOS (Batch A+ Z)		9-MOS (Batch A)	
	Assignment using CTP			
Composite	M	SD	M	SD
10-test	.23723	.033	.22277	.026
5-test	.22149	.035	.20236	.027
	Assignment using SQT			
10-test	.24245	.026	.23904	.026
5-test	.22011	.030	.22015	.026

Note. Averages were computed across fifteen replications of composites for the 15-MOS condition and ten replications of composites for the 9-MOS condition. MOS = Military Occupational Specialty; CTP = core technical proficiency; SQT = skill qualification test; 10-test = 10-test full least squares estimates composites; 5-test = 5-test best weighted composites.

Second, we examine the results for the 9-MOS condition where the CTP criterion measures include both equally weighted hands-on and skill proficiency components. In the results for the 15-MOS condition noted earlier, the CTP measure for 6 Batch Z jobs have only the skill proficiency component. For the 10-test composites, the MPP is .223 using CTP and .239 using SQT. For the 5-test composites, the MPP is .202 and .220. The differences between the two criteria are .016 and .018 of a standard deviation unit respectively. The results are comparable, but the SQT criterion yields larger MPP than does the CTP criterion when the predictors are ASVAB tests selectively augmented by AO. These

results show the general comparability of outcomes obtained using either criterion, although the magnitude of MPP consistently favors the SQT criterion.

E. Stability of MPP

Table 9 is intended to show the stability of results obtained using either of the criteria in different situations. In this comparison, we substitute in the analysis sample weights obtained from other data sets in place of the weights obtained in the present analysis.

Table 9

Mean Predicted Performance for ASVAB Assignment Composites Using "Substitute" Weights for 15 MOS

	Assignment using CTP		Assignment using SQT	
Composite	M	SD	M	SD
9-test	.11730	.026	.14702	.022
5-test	.11382	.027	.13946	.026

Note. ASVAB = Armed Services Vocational Aptitude Battery; MOS = Military Occupational Specialty; CTP = core technical proficiency; SQT = skill qualification test; 9-test = 9 test full least squares estimates composites; 5-test = 5 best weighted composites.

The samples used in determining the substitute weights were entirely independent of the samples used in the present analysis. For the CTP criterion, we used the 1983-84 concurrent validation data set of Project A. For the SQT criterion, we used FY 1988-89 "cohorts" of the sample included in the present analysis. The SQT and ASVAB scores were obtained from ARI's data files. Because ARI's data set contained only ASVAB test scores, the experimental AO test was dropped from this analysis and comparisons for the two criteria were made for the 9-test and 5-test composites.

First examining the results for the 9-test composites across the 15 MOS, the MPP after optimal assignment is .117 using CTP and .147 using SQT. Examining the results for the 5-test composites, the MPP is .114 using CTP and .139 using SQT. In the case of using substitute weights, an indicator of expected stability, the SQT criterion provides statistically significant higher MPP at the .05 level, the MPP differences between the two criteria being .030 for the 9-test composites and .026 for the 5-test composites.

It should be noted that this substitute analysis of weights is a very rigorous test of stability. The comparison uses two independent samples from different time periods in computing MPP. Additionally, several of the jobs in the job families are different, and for both the CTP and SQT criteria used in the substitute analysis there were slight modifications in measurement methods and content compared with

measurement methods and content used for the criteria in the longitudinal data set shown in Table 8. Most importantly, to compute stable family weights for operational use, sample sizes of several thousand are required rather than sizes in the hundreds that were used for some job families in the longitudinal analysis. The reduced magnitude of MPP is also due, in part, to the use of partially biased MPP (Table 8) with totally unbiased MPP (Table 9). Thus the results obtained in the stability analysis shows considerably more shrinkage than in traditional cross validation studies. Nevertheless, keeping in mind that our primary focus is on the substitutability of the two criteria, the stability analysis findings indicate significantly greater stability of the SQT over CTP against ASVAB predictors.

It is important to note that even small increments in MPP standard scores have been shown to translate into significant and practical estimates of dollar gain. A completely unbiased estimate of MPP using the full least squares estimate model of optimal assignment, showed an MPP of .17 compared to an MPP of .05 for the operational aptitude area composites used by the Army (Johnson & Zeidner, report in preparation). Similar gains over the operational aptitude area system were shown by Scholarios, Johnson and Zeidner (1994) and by Nord and Schmitz (1991) each using different independent data sets than were used by Johnson and Zeidner(report in preparation). Nord and Schmitz estimated the net economic value and cost of .12 gain in MPP they found in their study for the full LSE composites over the operational aptitude area composites. They used both a net present value model of performance valuation (Brogden, 1951; Schmidt & Hunter, 1983) and a more traditional opportunity cost model. The opportunity cost approach was used because of the subjectivity and possible unreliability of net present value and because of concern related to net present value procedures being applied to public sector activities in which no clear valuation of output is possible. Under both models, MPP standard score gains of less than 0.1 were shown to imply significant and practical gains from an improved system of selection and classification.

Another indication of the utility of small differences in MPP is the finding that optimal classification provides about twice as much gain in predicted performance as does gain from selection alone, i.e. the gain from classification is about as large as the gain from selection (Johnson & Zeidner, report in preparation; Johnson, Zeidner & Leaman, 1992; Nord & Schmitz, 1991; Scholarios, Johnson and Zeidner, 1991; Zeidner & Johnson, 1994).

F. Factor Structure

From an examination of the results of principal components analysis yielding from 3 to 9 factors, we judged a six rotated oblique factor solution as providing the best simple structure. Table 10 provides a comparison of CTP and SQT coefficients (loadings) for six rotated oblique factors in the joint predictor-criterion space for the 10-test battery. The covariance matrix was composed of 30 predicted

performance variables, 15 MOS using CTP and the same 15 MOS using SQT. The actual criteria were not factored, rather, the predicted criteria, or LSEs, were used. Table 11 shows the intercorrelations among the six oblique factors.

Table 10

Comparison of CTP and SQT on Coefficients for Six Rotated Oblique Factors in the Joint Space (10-test)

MOS	Criterion	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
11B	CTP	-.0311	**.4023**	.0643	.1753	.1473	.0105
	SQT	.1965	**.3271**	.0149	-.0017	.0039	-.0371
12B	CTP	-.0213	.2277	**.3160**	.2157	.0882	.0826
	SQT	.0247	.1108	**.1638**	.0589	.1521	-.0265
13B	CTP	**.1037**	**.2636**	**.0174**	**.0035**	.1203	**-.0099**
	SQT	.1206	-.0512	**.1835**	.1766	.0720	.0031
16S	CTP	.1606	.0928	.0968	.0153	**.2910**	.0226
	SQT	.2144	.1597	.0596	-.0558	**.2486**	.0331
19K	CTP	.0284	**.5145**	-.0235	.0878	.0934	-.0321
	SQT	.1712	.1573	**.2369**	.1685	.0448	-.0458
31C	CTP	-.0414	-.0297	**.5220**	-.0222	.0308	.0574
	SQT	.1667	-.0242	**.3532**	.2206	.0221	-.0922
54B	CTP	.0134	**.3223**	-.0519	.1822	.0633	.0607
	SQT	.1162	**.4900**	.1945	-.0120	-.0508	-.0186
55B	CTP	.1651	.1624	.1188	-.0010	.0606	**.3418**
	SQT	**.3298**	.0464	-.0654	.0767	-.0219	.2675
63B	CTP	.0454	**.3248**	.2104	.0115	-.0133	.1299
	SQT	-.0195	**.3975**	.3974	.0032	.0068	.0205
71L	CTP	**.5041**	.0327	.0582	.1257	.0214	.0550
	SQT	**.5249**	.0379	.0051	-.0389	-.0015	.0779
76Y	CTP	**.3064**	.0727	.0029	.1353	.1019	.0314
	SQT	**.3473**	.0667	.0680	.0186	.1165	-.0373
88M	CTP	.1387	**.3386**	.0219	.0856	.0584	.0367
	SQT	.0519	.1232	.0986	**.1702**	.1583	-.0011
91A	CTP	**.2452**	.2401	-.0558	.1525	.0583	.0279
	SQT	.1886	**.3341**	.0405	.0428	.0189	-.0136
94B	CTP	.0811	.1012	.1051	**.4768**	-.0447	-.0526
	SQT	.0919	.2416	.0674	**.2969**	-.0788	.0802
95B	CTP	.0197	.2287	.0041	**.5265**	.0367	.0718
	SQT	.1589	.1540	.0642	**.2615**	.0861	.0577

Note. CTP coefficient listed first, SQT second for each factor. Bold indicates highest coefficient for each MOS on a specific factor. See table 5 for explanation of job (MOS) abbreviations.

The results show that 10 of the 15 pairs of MOS have their highest coefficients (loadings) on the same factor for both CTP and SQT. For the remaining 5 MOS, the highest loading on a factor for one criterion is the second or third highest loading on that same factor for the other criterion. The

intercorrelations among the 6 factors range from .301 to .710, with a mean intercorrelation of .507. Comparison of the factor loadings across the 6 factors for the two criteria for each of the 15 MOS shows magnitudes and patterns of loadings that are quite similar.

Table 11

Intercorrelations Among Six Oblique Factors in the Joint Space

Factor	1	2	3	4	5	6
1	1.0000					
2	.5778	1.0000				
3	.4983	.6984	1.0000			
4	.6076	.7096	.5989	1.0000		
5	.4044	.4913	.3884	.5380	1.0000	
6	.3293	.5420	.3006	.5009	.4198	1.0000

It is interesting to note that similar results in other studies show nontrivial multidimensionality of the joint predictor-criterion space and its usefulness in classification and assignment (Statman, 1992; Zeidner & Johnson, 1994). The results in the present study are completely dependent on the 10 cognitive predictors (ASVAB plus AO test). The factor structure and loadings would vary as the composition of variables change. However, the objective of the principal component analysis was to investigate the factor structure and pattern of loadings of the two criteria as defined by the 10 predictor tests, since the overall goal of the study was to determine the substitutability of one criterion for the other in the contest of ASVAB. The principal component analysis contributes to understanding the construct validity of the criteria in making classification decisions concerning ASVAB.

G. Reliability

Table 12 presents reliabilities for the hands-on and job knowledge components of CTP and for the six MOS SQT reliabilities matching the paper-and-pencil job knowledge test components. The two paper-and-pencil criteria have mean reliabilities of .88 (job knowledge) and .83 (SQT), and the mean reliability of the hands-on component of the CTP has a reliability of .76. All validities shown in the table have been corrected for attenuation. The overall hands-on component of CTP is somewhat less reliable than the two paper-and-pencil measures.

Table 12

Comparison of Hands-On, Job Knowledge and SQT Corrected Reliabilities by MOS

Hands-On Tests[1]		Job Knowledge Tests[1]		SQT Tests[2]	
MOS	Reliability	MOS	Reliability	MOS	Reliability
11B	.85	12B	.80	12C	.82
13B	.83	16S	.90	16S	.77
19K	.60	54B	.85	54B	.79
31C	.88	55B	.93	55B	.88
63B	.78	76Y	.87	76Y	.89
71L	.82	94B	.87	94B	.82
88M	.75				
91A	.85				
95B	.51				
Mean	.76		.88		.83

[1] Split-half reliabilities
[2] Alpha reliabilities

H. Job Elements

In summary, the hands-on component of the CTP was developed for task elements identified by subject matter experts (SME) using task analysis information from a number of manuals. The job knowledge component of CTP was also developed essentially using the same task information as was used for hands-on tests. Task elements composing the operational SQT measures were developed by SME of the proponent agency for the MOS, again using similar information as was used for CTP. We judge, then, from the descriptions of the rigorous criteria development sequences employed, that there was the opportunity of considerable overlap of task elements in the development of criteria employed for a given MOS. This judgment is not based on a checklist of task elements measured by each criterion. Rather it is a global judgment of overlap because of the great similarity in general approach and in the sequence of steps employed in defining criterion content and method of measurement.

Summary and Conclusions

A. Summary

The major goal of this research was to determine if SQT criterion measures could be substituted for hands-on CTP criterion measures for purposes of making decisions concerning the development and evaluation of ASVAB classification procedures. Some researchers consider that two criterion measures are substitutable when their estimated utilities are the same even though many particular dimensions of those measures may differ. Predictor-dependent methods, such as used in the principal component analysis in the present study do not provide broad-based evidence of factorial structure and construct validity of the two criterion measures. But predictor-dependent methods are supplements to other indices used in this study to determine the relative utility of two measures in the context of ASVAB.

The overall question concerning similarity raised in the present study was: Would similar decisions be made using either the CTP or the SQT? And if the decisions differed somewhat: Would the results be considered equivalent? Seven indices of criterion similarity were evaluated.

First, the relationship between the CTP and SQT measures after corrections for attenuation was determined. The direct correlations between the total scores of the two criterion measures is .46 and the correlation between the two predicted performance scores is .92 for the 9-MOS condition. Thus, the two measures were highly correlated in the joint predictor-criterion space or in the valid space. These findings suggest that the two criteria measure different dimensions, but when only those components of the criteria overlapping the predictor are considered--the valid or predicted performance space-- either criterion could serve as a surrogate for the other in making classification decisions concerning ASVAB.

Second, in considering the tests selected for the best 5-test composites for each of the 8 job families, 26 of 40 pairs of comparisons select the identical set of tests for the best 5-test composites using either CTP or SQT. Further the mean multiple correlation for CTP is .57 and for SQT is .54. Thus, when the test selected for the best 5-test composite for a job family differed by one or two tests, depending on the criterion used, the overall predictive validity results are quite similar. However, for two job families, practical differences in predictive validity are found. These findings suggest that in the critical decision area of choosing which ASVAB tests to use in an assignment composite for a job family, nearly the same tests are chosen using either criteria. If the tests are found to differ to some extent, depending on the criterion used, the overall multiple correlation values across all jobs are about the same. The multiple *R*, as used in this situation, is analogous to selection results assuming independent pools of

applications. Again these results are supportive of using either criterion as a surrogate for making similar ASVAB decisions and for obtaining comparable results.

Third, keeping in mind that the focus of this study is on finding a surrogate criterion in the context of classification efficiency, the overall MPP for the 9-MOS condition is .22 using CTP and .24 using SQT, a practical and statistically significant difference at the .05 level. Three of the job families, however, have MPP that vary by the criterion measure utilized.

We conclude from this finding that overall the operational SQT is at the very least as promising as CTP when evaluating test composites from among Army ASVAB tests and AO, an experimental cognitive test. One possible explanation for the results obtained using SQT tests as criteria may be that they and the cognitive predictors used in this study are multiple-choice paper-and-pencil tests and therefore MPP may be inflated because of common method variance. However, as mentioned above, both criterion measures essentially select the same tests for each composite and use comparable weights of the tests in the composites.

Fourth, in a rigorous test of stability using substitute weights obtained from different samples and from several different jobs composing job families, the MPP is found to be .117 using CTP and .147 using SQT in the 15-MOS condition, a practical and statistically significant difference at the .05 level. Thus the SQT provides more stable results than does CTP using cross samples in which the data sets were quite different.

Fifth, in conducting a principal component analysis of CTP and SQT in the joint predictor-criterion space, we find that six rotated oblique factors provides good simple structure. We found that 10 of the 15 pairs of MOS have their highest coefficients (loadings) on the same factor for both CTP and SQT. We conclude that the factor structure and the loading across factors are quite similar for the two criteria, while recognizing that the results are predictor-dependent.

Sixth, the results show that the different criteria, that is, hands-on, job knowledge, and operational SQT have mean corrected reliabilities of .76, .88, and .83, respectively. In addition to having somewhat lower overall reliabilities, hands-on tests can lose reliability over time in field testing.

Seventh, from the description in the literature of the rigorous sequence followed in the development of each criterion component, we make an overall judgment that there appears to be overlap in many of the task elements in each criterion, although we did not perform an independent analysis of the task elements covered in each criterion.

B. Conclusions

Hands-on performance measures are considered by many as benchmarks of job proficiency. Project A findings show that the core technical proficiency (CTP), measuring hands-on performance, is adequate for classification research. Wise, McHenry and Campbell (1990) showed that CTP measures require separate prediction equations across jobs and thus are sensitive to differential assignment methods. However, in Project A, hands-on criteria were developed for only 9 MOS. Accepting that hands-on criteria have the desired attributes of construct validity, reliability, and relevance, they remain prohibitively expensive to construct and administer. Classification research requires job specific criteria to adequately represent core jobs for more than 250 entry-level jobs. Consequently, a surrogate is necessary for the development and evaluation of classification procedures utilizing ASVAB.

In recent classification research the present investigators specified the use of least squares assignment composites based on the full set of ASVAB for 66 core job families. This research utilized ARI's data set of ASVAB and operational SQT scores. But before full consideration be given to proposed changes based on this research, the suitability of SQT as a criterion measure needs to be empirically determined.

The present study demonstrates the adequacy of job knowledge criteria and/or the operational SQT measures, to serve as surrogates for hands-on criteria, the CTP measures. The conclusion is based on the findings that nearly the same decisions are reached and that equivalent findings are obtained for both types of criteria in developing classification procedures for ASVAB. Research findings for seven indices showed that the use of either criterion resulted in: (1) similar tests being selected for assignment composites and receiving similar weights; (2) similar outcomes as measured by MPP, the measure of classification efficiency, and stability of MPP in cross samples using different data sets reliabilities of the two criteria; and (3) similar underlying dimensions as measured by intercorrelations between predicted performance scores of the two criteria, their factor structure and loadings; and overall judgment of overlap in task elements measured. The use of job knowledge performance measures, with comparable psychometric qualities of the SQT criteria used in the present study, should no longer be a central issue in classification research using ASVAB.

References

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology*; Vol. 2. Palo Alto, CA: Consulting Psychologists Press.

Brogden, H. E. & Taylor, E. K. (1950). The dollar criterion - Applying the cost accounting concept to criteria construction. *Personnel Psychology, 3,* 133-154.

Campbell, C. H., Ford, P., Rumsey, M. G., Pulakos, E. D., Borman, W. C., Felken, D. B., deVera, M. V. & Riegelhaupt, B. J. (1990). Development of multiple job performance measures in a representative sample of jobs. *Personnel Psychology, 43,* 277-300.

Campbell, J. P. & Zook, L. M. (1994, March). *Building and retaining the career force: New procedures for accessing and assigning Army Enlisted personnel.* Annual report, 1991 fiscal year. ARI Research Note 94-10. Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences. (AD A278 726)

Campbell, R. C. (1994, February). *The Army Skill Qualification Test (SQT) program: a synopsis.* Interim Report, IR-PRD-94-05. Alexandria, VA: Human Resources Research Organization.

Chronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Dunnette, M. D. (1963). A note on *the* criterion. *Journal of Applied Psychology, 61,* 206-209.

Gottfredson, L. S. (1991). The evaluation of alternative measures of job performance. In A. K. Wigdor & L. M. Green (Eds.), *Performance assessment for the workplace; Vol. 2: Technical issues.* Committee on the Performance of Military Personnel, National Research Council. Washington, D.C.: National Academy Press.

James, L. R. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin, 80,* 75-83.

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 classification efficiency by restructuring Army job families.* Technical Report 947, Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A250 139)

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A. & Ashworth, S. (1990). Project A validity results. The relationship between predictor and criteria domains. *Personnel Psychology, 43,* 335-354.

Nord, R. & Schmitz, E. (1991). Estimating performance and utility effects of alternative selection and classification policies. In J. Zeidner & C. D. Johnson, *The economic benefits of predicting job performance; Vol. 3: The gains of alternative policies.* New York: Praeger.

Project A (1990). Project A: The U.S. Army selection and classification project. *Personnel Psychology, Special Issue 43,* 231-378.

Schmidt, F. L. & Hunter, J. E. (1983). Individual differences in productivity. An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology, 68,* 407-414.

Scholarios, D. M., Johnson, C. D. & Zeidner, J. (1994). Selecting predictors for maximizing the classification efficiency of a battery. *Journal of Applied Psychology,* 412-424.

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness. The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally.

Statman, M. A. (1992, August). *Developing optimal predictor equations for differential job assignment and vocational counseling.* Paper presented at the American Psychological Association Annual Meeting, Washington, D.C.

Wherry, R. J. (1957). The past and future of criterion evaluation. *Personnel Psychology, 10,* 1-5.

Wherry, R. J., Ross, P. F. & Wolins, L. (1956). *A theoretical and empirical investigation of the relationships among measures of criterion equivalence.* NTIS No. AD727273. Research Foundation, Ohio State University, Columbus.

Wigdor, A. K. & Green, B. F. (1986). Assessing the performance of enlisted personnel: Evaluation of a joint-service research project. Committee on the Performance of Military Personnel, National Research Council, Washington, D.C.: National Academy Press.

Wigdor, A. K. & Green, B. F., Editors (1991). *Performance assessment for the workplace; Vol. 1.* Committee on the Performance of Military Personnel, National Research Council. Washington, D.C.: National Academy Press.

Wise, L. L., McHenry, J. & Campbell, J. P. (1950). Identifying optimal predictor composites and testing for generalizability across jobs and performance factors. *Personnel Psychology, 43*, 355-366.

Zeidner, J. & Johnson, C. D. (1994). 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.