Study
Note
2006-03

Evaluation of Alternative Aptitude Area (AA) Composites and Job Families for Army Classification: A Reply

Cecil D. Johnson and Joseph Zeidner
J. Zeidner and Associates

United States Army Research Institute for the Behavioral and Social Sciences

March 2006

Approved for public release; distribution is unlimited.

20060414067

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

A Directorate of the Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

MICHELLE SAMS
Technical Director

ZITA M. SIMUTIS
Director

Technical Review by

Peter M. Greenston, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Study Note 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: DAPE-ARI-MS, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

FINAL DISPOSITION: This Study Note 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 findings in this Study Note are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED (from. . . to)
March 2006	Final	January 2005 – December 2005

4. TITLE AND SUBTITLE Evaluation of Alternative Aptitude Area (AA) Composites and Job Families for Army Classification: A Reply	5a. CONTRACT OR GRANT NUMBER 5b. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Cecil D. Johnson and Joseph Zeidner	5c. PROJECT NUMBER 5d. TASK NUMBER 5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) J. Zeidner & Associates 5621 Old Chester Ct. Bethesda, MD 20814	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 2511 Jefferson Davis Highway Arlington, VA 22202-3926	10. MONITOR ACRONYM ARI 11. MONITOR REPORT NUMBER Study Note 2006-03

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Subject Matter POC: Peter M. Greenston

14. ABSTRACT *(Maximum 200 words)*: Differential assignment theory (DAT) and research findings bearing on initial personnel classification from a number of simulation experiments, based on large samples of Soldiers with both operational and experimental test scores and MOS specific performance scores, is drawn upon in a critical review of recent HumRRO research (ARI Study Report 2005-1). The report being reviewed recommends using nine job families with corresponding best-weighted test composites that have been corrected for restriction in range to the recruit population and then converted to Army standard scores. This is the second tier of the two tiered classification system (TTCS) proposed by Zeidner and Johnson. The HumRRO authors primarily examine differential validity and validity coefficients with standard errors to reach the conclusion that there is no need for the first tier of the TTCS. A number of issues on which there is disagreement with the HumRRO authors are discussed.

15. SUBJECT TERMS

Military personnel classification, Army Aptitude Area (AA) composites, Armed Services Vocational Aptitude Battery (ASVAB)

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	40	Ellen Kinzer Technical Publication Specialist 703/602-8047

Acknowledgement

The authors are deeply appreciative of the advice, technical comments, and editing assistance provided by Dr. Peter Greenston in the preparation of this report.

Preface

Joseph Zeidner, Cecil Johnson, and colleagues developed a proposed two-tier classification system for Army enlisted Soldiers. A recent ARI report (ARI Study Report 2005-1) documented an independent analysis and evaluation by scientists at the Human Resources Research Organization (HumRRO) of key components of this system. Zeidner and Johnson took issue with several findings / conclusions in this report and requested an opportunity to respond. This reply presents their response to the HumRRO evaluation. Although it is unusual to present a critique of one ARI report in another, it is ARI's view that this report elucidates key aspects of Zeidner and Johnson's Differential Assignment Theory and is valuable for that purpose. The points presented in this report are best understood if the reader has already read some of Zeidner, Johnson and colleagues' earlier ARI reports (1992, 1997, 2000, 2003a, 2003b) on Differential Assignment Theory and the ARI Study Report in which HumRRO's evaluation is presented (Diaz, Ingerick, & Lightfoot, December 2004).

Executive Summary

Differential assignment theory (DAT) and research findings bearing on initial personnel classification from a number of simulation experiments, based on large samples of Soldiers with both operational and experimental test scores and MOS specific performance scores, is drawn upon in a critical review of recent HumRRO research (ARI Study Report 2005-1). The report being reviewed recommends using nine job families with corresponding best-weighted test composites that have been corrected for restriction in range to the recruit population and then converted to Army standard scores. This is the second tier of the two tiered classification system (TTCS) proposed by the current investigators. The HumRRO authors primarily examine differential validity and validity coefficients with standard errors to reach the conclusion that there is no need for the first tier of the TTCS. A number of issues on which we disagree with the HumRRO authors are listed and briefly introduced below.

First, the HumRRO authors argue that the number of job families, the approach used in forming these families, and standardization vs. non-standardization of composite scores, should be evaluated in terms of effect on validity and differential validity coefficients and the standard error of these indices – in contrast to our reliance on comparison of the mean predicted performance (MPP) computed in independent cross samples to reflect the various experimental conditions.

Second, this reply rejects the HumRRO authors' recommendation to use job families identified by judgment and administrative considerations instead of clustering MOS using Horst's differential validity method to maximize the MPP obtainable by optimal assignment.

Third, this reply rejects the HumRRO authors' claim that it is essential to convert composite scores used for initial classification and assignment into Army standard scores in order to achieve a desirable quality distribution in critical MOS. We maintain that the desired quality distribution of MPP across MOS can be obtained by using least square estimates of the criterion as unstandardized scores to make optimal assignments while applying a quality constraint – thus achieving a higher average MPP and a maximum fit to the desired quality distribution.

Fourth, the HumRRO authors claim that the increased MPP that would result from the changing of job families from 9 judgmentally formed ones to a larger set of 20 to 40 optimally clustered MOS would provide little or no benefits. We claim that the benefits in terms of increased performance would be worth millions of dollars. Years of research experience tell us that the magnitude of these benefits obtainable from improving the personnel classification process could never be achieved by improving the selection battery with new and possibly more or more difficult-to-administer predictor tests.

We agree with the HumRRO authors that our triple cross-validation for removing inflation due to sampling error from estimates of MPP could be further improved. More stable estimates of MPP could be computed by using all of the 240,000 observation cases available to compute each set of MPPs. Also, another source of sampling error that is introduced by permitting overlap between the sample used to cluster MOS into families and both the analysis and evaluation samples could be eliminated. These research design issues are discussed in the reply.

Contents

Page

Introduction 1
The Two-Tiered Classification System (TTCS) 2
Differential Assignment Theory (DAT) and Its Application 4
    Triple Cross-Validation Design 6
    Clustering MOS into Job Families 8
    General Mental Ability 10
Inconsistencies with respect to TTCS and/or DAT 13
    GMA, Number of Job Families, and Homogeneity 13
    Role of GMA in Classification 14
    Use of Differential Validity 15
    First vs. Second Tier Job Families and the MPP Continuum 17
    MPP and Standardization 18
    Experimental Design Issues 21
    Valuation of Classification Gains 23
    Potential Contribution of Non-Cognitive Variables 24
Summary of Objections 25
Conclusions and Recommendations 27
References 31

Evaluation of Alternative Aptitude Area (AA) Composites and Job Families for Army Classification: A Reply

Introduction

This reply is a critique of a HumRRO report by Diaz, Ingerick, and Lightfoot, "Evaluation of Alternative Aptitude Area (AA) Composites and Job Families for Army Classification", ARI Study Report 2005-01 (December 2004).[1] In our view this report reflects a lack of understanding of aspects of several relevant topics including: (1) the two tiered classification system (TTCS) as proposed by Zeidner, Johnson, and colleagues; (2) the differential assignment theory (DAT) utilized in the design and evaluation of the TTCS; (3) the magnitude of the impact on classification efficiency of general mental aptitude (GMA) as compared to the remaining predictors of job performance; and (4) the extent of the decrease in classification efficiency when assignments are made by scores converted to Army standard scores. Critical empirical relationships among variables essential to the design of a TTCS are also discussed, including the necessity of using predictor scores that are least square estimates of each MOS criterion as the basis of making optimal assignments to jobs.

We start with a summary of TTCS and DAT topics, and then discuss the specific instances in which the indicated report ignores or misrepresents the objectives or characteristics of the TTCS or violates one or more principles of DAT. We then discuss where and how a number of HumRRO recommendations are inconsistent with DAT, followed by what should be

[1] In an earlier report (Diaz, Ingerick, and Lightfoot, August, 2004) the authors verified Zeidner et al. research results, including "best weights" for the seven operational ASVAB subtests used to form the nine Army composites which correspond to nine operational job families currently used in the initial assignment of new Soldiers.

done in the way of further research before even considering accepting certain HumRRO recommendations.

The Two Tiered Classification System (TTCS)

In a recent operational change to the Army personnel system, the integer-weighted test composites for the nine operational job families were replaced by empirically weighted composites, using all seven Armed Services Vocational Aptitude Battery (ASVAB) subtests for each family. These composites are corrected for restriction in range in the youth population and then converted to yield Army standard scores.

Benefits provided by this change lie primarily in the greater validity provided, in contrast to the validity provided by the three / four subtest integer-weighted composites. These benefits also apply to personnel selection, classification and assignment, to vocational counseling of new Soldiers, application of minimum cut scores, and other administrative purposes.

A major drawback still existing after this change is that the nine job families were formed entirely by judgment and administrative considerations. Thus, these best weighted composites cannot provide maximum classification efficiency (as measured by mean predicted performance, MPP), in comparison to what is possible when MOS are clustered into families using Horst's differential validity method (Horst, 1954). A second drawback stems from the reduction in classification efficiency that results from converting the composite weights to provide equal means and standard deviations (SDs) in the youth population across all nine composites. This also causes a major reduction in classification efficiency. And a third drawback follows from the use of only nine job families ($m = 9$) and corresponding composites. This results in a costly reduction in MPP as compared to what would be provided by a larger number of job families,

formed with optimal clustering techniques and corresponding best weighted composites. The Two-Tiered Classification System (TTCS) is designed to overcome these drawbacks.

The TTCS is designed for use in the classification of recruits from the Army input population that is created through use of AFQT quality marks, academic accomplishments, justice system records, etc. to select applicants from a youth population. Recommended assignments for initial classification and assignment of recruits are obtained by using 25 to 150 best-weighted test composites of seven ASVAB tests; these composites correspond to the same number of first tier job families. Recommended assignments are optimized to maximize the MPP of those for which assignments are being made with a constrained linear programming algorithm that would exactly fill all training seats if the recruits accept their recommended assignments. A revised set of quotas could be utilized to reflect those who reject their recommendations.

The first tier TTCS composite weights are corrected to the Army input population from which the sample of assignees is drawn. No purpose would be served by standardizing first tier composites or composite weights to the youth population, and thereby requiring all composites to have means of 100 and standard deviations of 20 (or requiring all weights to be positive) since these composites are invisible to counselors, administrative personnel and examinees alike in the classification and assignment process. Moreover, MPP is considerably increased by correcting first tier weights for restriction in range to the Army input population, requiring each composite to have a mean of zero and standard deviation equal to the validity coefficient of each composite in the Army input population[2], and allowing negative weights.

[2] As contrasted with second tier composites which are converted to Army standard scores.

The selection process that yields an Army input population uses AFQT criteria to select Army applicants from the youth population, along with a consideration of academic achievement, health, and criminal behavior. The use of AFQT in the initial selection into the Army can be supplemented by using cut scores on nine second tier composites. The TTCS calls for the second tier composites (which have been standardized to the youth population, where each composite has a mean of 100 and a standard deviation of 20 and composite weights are required to be positive) to be used in a subsequent selection type process in which eligibility for each of the 150 first tier job families is determined by use of MOS specific cut scores on one of the nine second tier composites that corresponds to the job family which contains the MOS being considered. The scores for these nine second tier composites are visible to counselors, administrative personnel, and examinees and it is essential that the composite weights for the composites of this visible tier be positive only, have equal means and equal standard deviations for all nine composites. The second tier composites are intended for use by counselors, administrative personnel and by new Soldiers in self-evaluation of their own aptitudes for alternative military training and eligibility for Army school programs.

Differential Assignment Theory (DAT) and Its Application

Empirically Determined Relationships Among TTCS Variables

H. Brogden (1955) established that the most efficient test composites for assignment of personnel are least square estimates of predicted performance composites (PPs) of each job family criterion, and P. Horst (1954) provided a method for aggregating jobs into job families to

maximize the average differential validity for a set of job families, thus maximizing potential MPP for that set. Brogden (1959) provided a formula for estimating MPP as a function of the number of job families (m), the mean of the validity coefficients of the associated best weighted composites (R), and the mean of the inter-correlation coefficients among the composites (r): BMPP = $f(m)\, R\, (1 - r)^{1/2}$. The MPP maximized by Horst's optimal job clustering procedure can be expressed as BMPP when one makes the same assumptions as Brogden made to derive BMPP (Johnson & Zeidner, 1995).

The Brogden estimate of MPP pertains to an optimal assignment of N Soldiers to m job families using m assignment composites that are each LSEs of the m separate criterion variables. Brogden provides a table for $f(m)$ for values of m of 1 through 15. The variable R is the average of m correlation coefficients between LSEs and criterion scores, and r is the average of $m\,(m\text{-}1)$ inter-correlation coefficients among the LSEs. This formula provides a good approximation of empirically determined MPP for small values of m but provides an under-estimate of MPP for larger values of m. Both this formula and empirical results indicate that high values of MPP can be obtained for even high values of r.

Since R is the average of cross-validated validity coefficients, N (the size of the analysis sample) is positively correlated with the magnitude of R. For any set of jobs, the correlation between sample size and size of unbiased validity coefficients is positive (but low unless sample sizes are much lower than we used in evaluating the TTCS).

The value of r for a specified set of jobs decreases as the value of m increases for this fixed set of jobs. The three variables, r, R and m, are all predictors of MPP with a fairly complicated non-linear relationship between them. The comparison of the benefits between

alternative sets of job families must be in terms of MPP rather than in terms of *r*, *R* ,*m*, or size of analysis samples (*N*), evaluated separately or in pairs.

Empirical results for a set of 250 MOS initially clustered into 150 families (most of which are single MOS) show that MPP increases steadily as m increases to an unknown point between 66 and 85, then decreasing gradually out to the 150 family point. (Zeidner, Johnson, Vladimirsky & Weldon, August, 2000).

Triple Cross-Validation Design

An unbiased estimate of MPP for a designated set of assignees, tests, and job families can be computed through use of a triple cross-validation design. This design as used by Zeidner and colleagues consists of the following components or processes as described below.

A data set of 260,000 Soldiers where each Soldier's record contains MOS, ASVAB test scores, gender and race, and a Skill Qualifications Test (SQT) criterion score, is randomly divided into two samples, one containing 240,000 cases and a hold-out sample containing 20,000 cases. The 240,000 cases are first divided into separate MOS sub samples. Each of these sub-samples is then randomly divided into Sample A and Sample B, each containing 120,000 cases in total. The 20,000 cases provides a cross validation sample, Sample C, which is divided into 20 sub-samples without regard to the MOS assigned to each Soldier in the sample.

Sample A is used as an analysis sample, as a sub-sample separately for each MOS cluster (i.e., job family) in which regression weights are computed for the best weighted assignment composites associated with each job family. Each Sample A sub-sample has a parallel Sample B sub-sample.

The Sample B sub-samples specific to each job family are used as evaluation samples, in which regression weights for tests are computed to provide least square estimates of the MOS specific criterion within each job family.

As noted above, Sample C is a cross sample in which the regression weights computed in Sample A, or operational weights for one experimental condition, are applied to the Sample C test scores to produce assignment composite (AA) scores for each individual. As described below this sample is used to optimally assign individuals to job families and to subsequently compute a value of MPP for a specified set of job families, separately for each Sample C sub-sample of 1,000 cases. The comparable regression weights computed in Sample B are applied to these same Sample C test scores to produce predicted performance (PP) scores in each job family for each individual in Sample C.

The PP score for the job family to which each individual is tentatively assigned by the optimal assignment algorithm (constrained linear programming algorithm) is attached to that individual for the purpose of computing MPP. The mean across sub-samples of all such PP scores in each Sample C sub-sample is designated as the value of MPP for that set of job families in each sub-sample. The MPP values for each of the twenty Sample C sub-samples are averaged to provide the MPP, and the standard error of these estimates, for each set of job families within an experimental condition.

As noted above, a set of 240,000 Soldiers was divided into job family sub-samples and each such job family randomly divided into two equally sized sub-samples. Thus, one half of each job family was placed into Sample A and the other half into Sample B. Sample A and Sample B were used, respectively, as analysis and evaluation samples. The roles of Samples A and B were then reversed. As the roles of A and B are exchanged, C remains the cross sample.

Each MPP for all job family sub-samples is computed on a sub-sample of C. Thus there are two independent unbiased estimates of MPP that can be computed from reversing the roles of Samples A and B.

The average of the MPP computed for these two different roles of Samples A and B is the value of MPP for each job family in a data set. Each such estimate of MPP is unbiased in that the estimate of MPP is not inflated by back sample validity coefficients, "bouncing betas", or by non-zero inter-correlation coefficients among the tests used to compute composite weights.

Clustering MOS into Job Families

There are at least two widely different approaches for clustering MOS into job families that yield maximum MPP when the corresponding composites are used with a constrained LP algorithm (using MPP as the objective function) to make optimal assignments. The one we usually use produces families that maximize Horst's differential validity (Horst, 1954). The approach starts with the inter–r matrix among the total set of MOS, e.g., $m = 150$, and sequentially reduces m by one at each step by combining the two families whose merger minimizes the reduction in Horst's differential validity. We could go all the way to $m = 2$, with each set of families being optimal for maximizing MPP for that value of m, when LSEs are used to assign recruits to families using a LP algorithm. The second approach starts out by producing principal component (PC) factors, which are then rotated to simple structure and used to place MOS into optimal families corresponding to each rotated factor. Each MOS is placed in the family corresponding to the rotated factor for which its correlation coefficient is highest. This approach starts with the same inter-r matrix, with the integer one placed in the diagonals, to obtain as many principal component (PC) factors as there are tests available for use in the composites. With our current data this would be 9 tests (7 tests if the current operational battery

is used). These 9 PC factors would be rotated to obtain simple structure, and these 9 rotated factors correspond to the composites to be used to make optimal assignments to 9 families. The actual composites would be obtained as LSEs of these rotated factors based on the 9 tests as predictors. The 9 job families for use in making optimal assignments would consist of the MOS whose highest factor loading is on a particular rotated factor. Using the MOS in a particular family, a second round of PC factors could be computed, rotated into simple structure and up to 9 sub-families within each of the initial 9 families identified.

Either of the two clustering methods just described can be utilized to determine the relative benefits provided by different numbers of job families for a fixed test battery. The MPP obtained by a process that includes a triple-cross validation design takes into account the effect of sample sizes in computing validity coefficients and the varying inter-correlation coefficients among the tests. Inflation and bouncing beta effects due to sampling error are eliminated. The same approach, varied slightly, can be used to select tests from a larger set of experimental tests as a means of maximizing MPP, while eliminating the effect of sampling error introduced by optimal clustering in a back sample.

The 17 families proposed for use in generating an alternative set of second tier composites were obtained by using the first approach described above except that possible membership in these families was constrained to not cut across any of the 9 operational families. A similar constraint could be imposed using the MOS in each of the 9 operational families to compute inter-r matrices for the MOS LSE composites corresponding to each MOS in a particular operational family. PC factors could then be computed and rotated to simple structure and the 17 sets of MOS yielding the cleanest simple structure identified.

General Mental Ability

General mental ability (GMA) requires further definition before we can describe its role in an effective classification process. Possible definitions sometimes used for GMA include the following: (1) the entire set of cognitive tests in a test battery; (2) the largest principal component (PC) factor in a factor solution for a battery; (3) the Brogden "*g*" factor which by his definition fully determines the inter-correlation coefficients among the composites while contributing nothing towards classification efficiency; (4) the proverbial "*g*" factor that is valid against most performance criterion measures or academic grades, but contributes little to classification efficiency other than through a hierarchical classification process whose benefits are obtainable only by assigning the higher scoring individuals to the jobs whose corresponding composites are most valid. To evaluate the classification efficiency of GMA (using definitions 2 and 4), a PC factor solution of the MOS composites, yielding k factors, can be obtained. The k-1 smallest factors can be rotated into simple structure and the factor scores for each of these rotated factors used for optimal assignment. If there are k factors, these factor scores can be used to make optimal assignments to k job families that correspond to these k rotated factors. By the first definition, all of these factors may be GMA while only the first PC (largest) factor is GMA by definitions 2 and 4.

DAT includes a model of prediction variables that is used to optimally assign new Soldiers to job families. This model most definitely is not based on specific factors to obtain or to measure the benefits of optimal classification to the Army. Neither DAT nor Zeidner, Johnson, and colleagues are proponents of specific factor theory. The factor model of the predictor variables embedded in DAT is described in the following paragraphs.

DAT considers the impact that the use of a single general factor (GMA) or, in contrast, the impact a multiple (group) factor model of the predictors has on classification efficiency. The inter-correlation matrix among composites based on group factors has a rank equal to the number of tests (7) in the operational battery only when the number of families is 7 or less. These factors can be used as assignment composites. Transformed group factors may also be used as assignment composites with the number of composites ranging from 4 to 150. The matrix of inter-correlation coefficients among transformed group factors will be less than full rank when the number of composites exceeds 7.

The use of a single general factor score as the sole assignment composite can be used to optimally assign recruits to job families, resulting in what DAT refers to as hierarchical assignment. In hierarchical classification the individuals having the highest predictor scores (AAs) are assigned to the job family whose predictor composite has the highest validity coefficient, until the quota is filled. If the total set of recruits to be assigned is rank ordered on their "*g*" scores, and, after optimal assignment, the same set of recruits rank ordered on their PP scores based on these "*g*" scores, the rank order of each individual will be the same on both lists. This is because optimal assignment to maximize MPP can be accomplished by first rank ordering the families based on their average validity coefficients for "*g*". Assignments to the *m* families are then made in order of their average validity coefficients, starting with the most valid, and assigning the N_j unassigned recruits with the highest "*g*" scores to the MOS with the highest validity coefficients for "*g*" *(Nj* is the quota for the *j*-th family*)*. If an optimal assignment algorithm is used to maximize MPP when each PP is entirely based on "*g*", recruits with the highest aptitude are assigned to the high technology MOS, and recruits with lowest aptitude to

the combat arms MOS. Experience informs us that the prospect of making such assignments would alarm most personnel assignment, training and general officers.

As one goes from one to four factors (as long used by the Air Force) to make optimal assignments, the channeling of the higher aptitude recruits to the more highly technical MOS decreases, but produces more such channeling than when nine LSE composites are utilized in the classification process. While this channeling effect decreases considerably when 9 operational (second tier) composites are used to make assignments, it does not totally disappear until an even larger first tier battery is used.

Hubert Brogden's formula for estimating MPP after optimal assignment using least square composites (LSEs) as assignment variables was derived by making the assumption that validity coefficients of PPs could be described in terms of a factor model that included a "*g*" factor and a specific factor corresponding to each job family. Brogden's "*g*" factor was defined in such a way that it can make no contribution to classification efficiency, but does completely explain the inter-correlation coefficients (r) among the composites. The validity coefficients (R) are determined by the specific factors. All classification benefits are provided by these specific factors. While Brogden used a specific factor model to derive the formula for BMPP, he definitely did not believe such a model represented the personnel classification process, nor that BMPP should be used as more than an initial approximation of the classification efficiency provided by alternative sets of job families. Both Zeidner and Johnson were supervised by Brogden during a period in which they separately worked on research projects intended to increase classification efficiency. Both were continuously made aware of Brogden's views regarding research on classification, and both knew that he accepted a factor model that included group factors with high inter-correlation coefficients among the predictors. Brogden most

definitely did not rely on specific factors to form his views on personnel classification, even though such a claim is made by a number of "*g*" theory proponents, including the investigators who believe that there is little more than "*g*". Schmidt et al. (1988) emphasized that the inter-correlation coefficients among predictor composites could be above .95 and still bring about significant improvement to classification efficiency as a result of optimal assignment.

Inconsistencies With Respect to TTCS and/or DAT

GMA, Number of Job Families, and Homogeneity

It is well known to psychometricians working in the area of personnel classification research that optimal classification to a number of jobs can be accomplished using a single assignment composite. A measure of "*g*" is almost always the most effective such composite. This approach is often referred to as hierarchical classification because the highest scoring individuals are assigned to the job families yielding the highest validity coefficients for this single composite. A similar hierarchical effect may be present, but not necessarily dominant, even when several composites are utilized if either the "*g*" component is dominant in some or all of the composites and the validity of "*g*" varies across job families, or the validity coefficients of composites with respect to their corresponding job families vary across job families. It seems obvious that the smaller (larger) the number of job families, the more (fewer) MOS contained in each family, and the more (less) likely that "*g*" will dominate some or all of the composites. It is also clear that standardizing composites (i.e., converting them to have equal means and standard deviations in a reference population) should reduce hierarchical effects.

Role of GMA in Classification

The least credible position taken by the HumRRO report authors occurs when the authors state that the decline in MPP as the number of job families (m) approaches 150 indicates that differential validity "is attributable to GMA, not specific abilities and aptitudes" (p.28). The authors also claim that the increase in MPP as m increases (up to a point) is attributable to higher validity coefficients and that this trend is "largely a function of GMA". This statement ignores the fact that GMA as measured by the first principal factor makes a smaller contribution to classification efficiency than do each of the next two largest factors in a principal factor analysis factor solution. As noted earlier, the "*g*" factor as defined by Brogden makes no contribution to MPP. Also, it is clear that the percentage of valid variance due to prediction by GMA decreases as the number of job families in a configuration (set) increases -- contrary to the statement by the HumRRO authors cited above.

The HumRRO authors (p. 28) make an argument we believe can be clarified using the following example. It is true that increasing the number of job families by 5 when starting from a set of 10 families (m = 10 to m = 15) provides considerably more increment in MPP than when starting with a set of 40 families and increasing to 45. The HumRRO authors argue that this is a clear indication "that a substantial portion of the differential validity among jobs is attributable to GMA, not specific abilities and aptitudes (p 28)". We maintain that the correct explanation of this decreasing benefit, resulting from further increasing m after m reaches approximately 60, is primarily the smaller sample sizes used to compute regression weights for the composites associated with each family as m is increased. Thus, while MPP obtained from optimally assigning individuals to m families always increases, for a specified value of R, as m is increased, increasing m adds to validity shrinkage as measured in the independent cross sample. As m

approaches the number of MOS having sample sizes that can support the computing of validity coefficients, validity shrinkage has a greater negative effect on MPP than the positive effect of increasing *m*.

A related explanation is provided by Brogden whose estimate of MPP, referred to as BMPP, has a term f(*m*) which increases BMPP as it increases. This term is asymptotic to a value which is closely approximated when *m* is 20 or larger. Thus, further increases in f(*m*) have virtually no effect on the value of BMPP. Intuitively, the average inter-correlation coefficient among composites (*r*) gets smaller and the average validity coefficient (*R*) gets larger as *m* increases. Both of these two trends, for *r* and *R*, respectively would intuitively increase MPP—if it were not for the increased shrinkage of *R* resulting from the smaller *N*s used to compute regression weights as *m* is increased.

Use of Differential Validity

The HumRRO authors use "differential validity" as an index of classification efficiency throughout their report and appear to believe that such an index is a reasonable substitute for MPP as a measure of classification efficiency. As noted in the above DAT section, Horst's index of differential validity is equivalent to Brogden's estimate of MPP (i.e., BMPP) if, and only if, Brogden's assumptions made in order to derive this formula are met. These assumptions include: (1) the full explanation of validity by specific factors unique to each predictor (composite), and (2) predictor inter-correlation coefficients are a function of a Brogden "*g*" factor that completely accounts for these predictor inter-correlation coefficients while having a zero correlation with the criterion variables. Neither Brogden nor we ever believed that differential validity is an adequate substitute for MPP as a means of comparing classification benefits provided by two different sets (configurations) of job families. However, we find

Horst's differential validity index useful in clustering jobs into job families and Brogden's BMPP index convenient in making preliminary estimates of benefits. It appears that the HumRRO authors' reliance on the concept of differential validity has led them to unduly focus on both the magnitude and standard errors of composite validity coefficients, rather than on MPPs, except for near the end of their report where they describe a simulation experiment that compared values and the standard errors of MPP. This focus on *R* has blurred the distinction between classification and selection benefits in the HumRRO report discussion.

The use of ten Sample C sub-samples to compute standard deviations of the MPP estimates provides a means of testing the statistical significance of the difference between the MPP values provided by two alternative sets of job family sets (e.g., 9 vs. 66). Such a test would provide the statistical significance of the differences in classification efficiency provided by two alternative approaches. We believe this test of classification efficiency is superior to that which is provided by a statistical test of the differences in the differential validity or validity coefficients.

It should be clear that the standard error of unbiased estimates of validity coefficients (between a best weighted composite and a criterion variable) where the regression weights are computed in the same sample as are the correlation coefficients is quite different from the standard error of unbiased validity coefficients where the validity coefficients are computed in a separate sample from the sample where the regression weights were computed. When the individuals used to compute the correlation coefficient have been placed in a sub-sample which uses a specified composite through use of an optimal assignment algorithm, a standard error computed as a function of sample size and size of the validity coefficient is an even poorer estimate of an SE that pertains to the MPPs used to measure classification efficiency.

First vs. Second Tier Job Families and the MPP Continuum

It should be noted that the sets of 9 and 17 job families and corresponding test composites are intended only for the second tier and there is little reason for either of these batteries of test composites to be compared with first tier sets of job families and corresponding test composites in terms of their classification efficiency. Thus, there is no logical reason to compare in terms of MPP either the 9 operational job families or the 17 job families derived from these 9 with the 40, 66 or 150 optimally clustered sets of job families. The first tier composites and job families are designed to achieve a radically different kind of benefit, i.e., classification efficiency, than the second tier composites and job families.

The first tier job family configurations of sets of 9 and 17 job families is unfortunately treated by the HumRRO authors as two points on a continuum consisting of the effect on MPP of the number of composites (m) in job configurations extending from $m = 9$, through $m = 17$ to $m = 150$. In fact, the $m = 17$ configuration described in the HumRRO report is a shredding out of the $m = 9$ configuration. The $m = 9$ configuration is based entirely on judgment that included administrative and military policy considerations. Each of the $m = 9$ job families is equal to the aggregation of job families of selected $m = 17$ configuration job families. When $m = 9$ configurations and $m = 17$ configurations are formed by using the optimal clustering algorithm provided by Horst (1954), the differences in MPP between these two optimally clustered configurations of $m = 9$ and $m = 17$ are considerably greater, as well as having higher values of MPP for both configurations, as compared to families (with $m = 9$ and $m = 17$) obtained by judgment. It is unfortunate that HumRRO did not choose to examine the effect on MPP of using optimally clustered $m = 9$, $m = 40, 66, 80$ (or some other midway values of m), and $m = 150$ to show the effect of size of m on MPP. As noted above, the $m = 9$ and $m = 17$ configurations

examined by the HumRRO authors were intended for possible use in the second tier and thus never recommended by Zeidner and Johnson for use in the classification process. Because of the configurations the HumRRO authors chose to examine, the evidence provided by the HumRRO authors does not elucidate the impact of m on MPP.

The continuum of MPP as affected by increasing the value of "*m*" when job families are optimally clustered using Horst's technique, and non-standardized, has been determined as follows: Starting with $m = 4$ paired with MPP = .1207 we can describe a continuum of pairs of *m* and MPP representing f(*m*) = MPP as follows: f(4) = .1207; f(6) = .1387; f(9) =:1662; f(11) = .1758; f(13) = .1720; f(15) = .1882; f(17) = .1931; f(21) = .1979; f(25) = .2032; f(40) = .2118; f(66) = .2120; f(85) = .2104; f(104) = .2068; f(127) = .2025; f(150) = .195 (Zeidner, et al., 2000). These MPP values are obtained from PPs scaled to provide a zero value of MPP in the Army input population. It is interesting to note that in the same study, the MPP provided by the selection process using PPs with a mean of zero in the youth population was .167. Thus the values of MPP using PPs scaled to have means of zero in the youth population (in contrast to the use of the input population to scale PPs used in computing MPPs for measuring classification efficiency) are equal to the above values of MPP scaled to the Army input population plus .167. The difference in MPP due to classification between $m = 9$ and $m = 66$ was shown to be slightly more than one-fourth of the MPP provided by selection.

MPP and Standardization

When discussing the use of standard scores, we are referring to a two step process in which PP scores which have been corrected to the Army input population are corrected to statistical standard scores (SSS) that have a mean of zero and an SD of 1.0 in the Army input population: SSS = $(\text{SD of PP})_j / R_j$, where R_j is equal to the validity coefficient for the jth job

family; and followed by computation of an Army standard score (SS), which has a mean (M) of 100 and an SD of 20 in the youth population, for the jth job family: $(SS)_j = [\ (SSS)_j + A_j\ ] / B_j$, where A_j and B_j are constants for each job family and are selected to correct for both restriction in range due to the selection process and the change of scale from M = 0 to M = 100 and from SD = 1 to SD = 20. One notes that R_j is smaller for combat arms and other MOS for which performance is less dependent on GMA—such as MOS for which performance is predicted by mechanical, other vocational or gaming skills that are measured by group or specific factors. The standardization process that converts from PP scores to SSS scores for a job family composite reduces MPP more for MOS primarily predicted by GMA than for other MOS. The second step, standardizing from SSS to SS, inserts noise through use of the A_j and B_j constants that further reduces MPP. Reduction in MPP due to both steps is greater for MOS for which selection variables such as AFQT, high school graduation, second tier cut scores and police records have a higher correlation coefficient with the MOS performance criterion variable. Intuitively, standardizing composites reduces the total MPP while also increasing MPP for most combat arms MOS at the expense of decreasing MPP for the other MOS.

The HumRRO authors make the point that, "MPP estimates tended to be systematically lower when weights were standardized," and that, "evidence for the significant interaction between standardization and job configuration can be seen in that differences in MPP by standardization increased as the number of jobs increased." We have noted this same relationship for more than a decade but completely disagree with the HumRRO authors as to the operational implications of this relationship. We believe that this indicates the desirability of not converting PPs to a scale which provides equal SDs when classification efficiency is the sole objective (i.e., as is true of first tier composites). We agree that it is appropriate to standardize

composites to have equal SDs when the primary objective is to determine whether a minimum cut score is met by a recruit or applicant for assignment to a specific job (i.e., as is true of second tier composites).

While we believe that standardizing composites is desirable for second tier composites, we believe, contrary to the position taken by the HumRRO authors, that standardizing the first tier composites is undesirable because of the resulting decrease in MPP. The HumRRO authors argue that this decrease should be accepted as the price of obtaining an increase in MPP in certain critical MOS. However, this latter benefit could probably be obtained with a much smaller reduction in MPP by constraining expected MPP in these critical MOS to be above a specified value during the optimal assignment process in Sample C in order to compute MPP. This doubly constrained process, to meet quality standards and quotas, could then be duplicated in the optimal assignment process for providing operational assignment recommendations by an algorithm that duplicates this effect for each set of recruits.

The HumRRO report authors argue that the increase in MPP provided by the first tier composites as m approaches 150 would be lost by operational efforts to maintain desired quality standards in those MOS that yield lower validity coefficients (e.g., combat arms and other relatively non-technical MOS). We have previously noted that MOS or job families with lower validity coefficients are likely to have lower MPP and we agree that this is undesirable, but we believe the solution proposed by the HumRRO authors is akin to throwing the baby out with the bath water.

It is not known at this time whether the higher MPP obtained from using non-standardized composites and/or a larger value of m would be negated by operational decisions; this is one of the issues we hoped to clarify in a field test. Since all recommended assignments are expected to

be implemented only if not rejected by the recruit or modified for operational reasons, we had not included the additional complication of constraining optimal assignments to provide a desired quality distribution across MOS. However, rather than following the HumRRO approach to meeting operational constraints (without empirical evidence), we propose the conduct of a simulation experiment in which optimal assignments are constrained to meet the desired quality distribution using *m* equal to 9, 66, 80 and 150. This additional constraint could be accomplished by adding constants to the PP scores that would raise the mean PP scores in critical MOS to a desired minimum level during the simulated optimal assignment process using Sample C sub-samples. We believe the first tier would provide a significantly higher total MPP value for job families with *m* between 40 and 85, as compared to the MPP provided by the second tier job families, when the optimal assignment algorithm includes this additional constraint.

Experimental Design Issues

The psychometric literature commonly defines the double cross-validation design in terms of two samples drawn from the same universe (e.g., A and B) in which best weights are computed for one sample (A) and applied to the scores of a second sample (B) to compute LSEs that are then validated in that second sample. The triple cross-validation design, as used by Johnson and Zeidner and other practitioners of DAT for more than a decade to compute unbiased estimates of MPP, utilizes a third independent sample (C) to accomplish the optimal assignment process and to then compute MPPs. Samples A and B provide two independent sets of regression weights that are used to compute pairs of LSEs of the criterion, for each job family for use in Sample C. One of these pairs of LSEs is designated as the analysis LSE (e.g., weights computed in Sample A) and the other is designated as the evaluation LSE (e.g., weights computed in Sample B). The optimal assignments are made in Sample C using the analysis

LSEs and the evaluation PPs are used in Sample C to compute MPP values for each job family. The roles of Samples A and B in computing either analysis or evaluation PPs is then reversed, providing two distinct estimates of MPP that are averaged to provide the MPP for each job family. On page 22 the HumRRO authors imply that Zeidner, Johnson, and Associates have, until recently, used the traditional double cross validation design to compute values of MPP, when in fact only the triple cross-validation design has ever been used by Zeidner and Johnson when using actual data in research to determine classification efficiency, in contrast to using synthetic scores randomly generated from estimates of population values. The HumRRO authors correctly describe the triple cross validation design on pages 23-24 but incorrectly refer to it as a double cross-validation design.

We agree with the HumRRO authors (pp. 14-16) that the use of data to optimally cluster MOS into job families that is not independent of the data used to compute MPP values will, to an unknown extent, inflate the positive relationship between the number of job families (m) used to compute MPP and the magnitude of MPP. A simple modification of the triple cross-validation design would completely eliminate this source of sampling error in both the clustering of job families and the computing of regression weights. Both sources of this sampling error can be eliminated by using Sample A to cluster MOS into families and compute regression weights for AAs, while Sample B is used to compute regression weights for PPs. The roles for Samples A and Sample B are then reversed.

The HumRRO report authors (pp. 23-24) describe a design that is similar to the DAT triple cross-validation design but differs in that their Sample C contains 49 sub-samples of 5,000 instead of 20 sub-samples of 1,000 cases. Each of their 49 Sample C sub-samples is used in conjunction with one half of the data that remains after the removal of each Sample C sub-

sample. One half of each such data pair is used as an analysis sample to compute regression weights for the AA composites and the other half as an evaluation sample to compute regression weights for the corresponding PP composites, with both the AA and PP composites used in the Sample C sub-sample to make optimal assignments and compute MPPs for each job family. Each of HumRRO's 49 estimates of MPP are computed using highly overlapping analysis and evaluation samples to compute regression weights for the 49 different estimates of MPP.

In this type of HumRRO design, MPP is computed as the average of 5 Sample C sub-samples with 1,000 cases in each of the 49 sets of sub-sample to create 49 separate (but far from independent) estimates of MPP. In the Zeidner and Johnson design, MPP is computed as the average of 20 Sample C sub-samples using only two independent estimates of regression weights for each configuration, providing an estimate of MPP for each configuration based on the average of 40 computed values of MPP, whereas the HumRRO design provides for computing 49 estimates of MPP to be averaged into a single estimate of MPP for each configuration. The standard deviations of the MPP estimates for each configuration involve using two independent sets of regression weights in the Zeidner and Johnson design, in contrast to the 49 different but overlapping samples for computing regression weights in the HumRRO design. We agree that the HumRRO design makes a better use of the total data set to provide a more stable estimate of MPP for each configuration, but does not provide as good an estimate of the standard error of MPP as does the Zeidner and Johnson design.

Valuation of Classification Gains

In the HumRRO report (p. 32), gains in MPP from using the $m = 17$ or $m = 150$ configurations are dismissed as follows: "...differences in mean predicted performance (MPP) among the 9, 17, and 150 AA test composites were not practically significant, especially after

taking into account estimation error in MPP." Nord and Schmitz (1999) argue, using the "opportunity cost approach" widely accepted by economists, that two Army standard score points improvement in the mean score for the AA composite corresponding to the MOS to which each recruit is assigned, is worth over 600 million dollars using a 1984 data set. A set of optimally assigned full least square composites for 9 job families was contrasted to the existing integer weighted composites as the baseline. Using the opportunity cost method the gains in predicted performance and costs of recruitment, retention, and training are compared for the two systems to estimate what it would cost to obtain the mean predicted performance obtained by LSEs of composites which are LSEs of performance criteria by raising the selection standards while retaining the then current AAs. It is hard to see how the gains in MPP provided by the m = 150 configuration, in contrast to those provided by the $m = 9$ configuration, can be so readily dismissed by the HumRRO authors.

Potential Contribution of Non-Cognitive Variables

The HumRRO authors claim (p.34) that, "...non-cognitive variables, specifically personality and vocational interests, could greatly extend the classification potential of cognitively-based composites". Zeidner and Associates have used the clustering technique of Horst to sequentially select tests from an experimental battery so as to maximize differential prediction. In one study (Johnson, Zeidner, and Scholarios, 1990), the best 10 of the 29 Project A predictors were sequentially selected to maximize H_d (Horst's index of differential validity). The best three tests for classification efficiency were cognitive tests and the fourth best test was an interest test. The selection order of the five selected cognitive tests was 1, 2, 3, 5, 8, and the selection order of the four selected interest tests was 4, 6, 7, and 9. The selection order of the single selected "perceptual-psychomotor" test was 10. In contrast, for an ARI experimental

battery developed earlier, the non-cognitive tests were selected (using a thumbnail index of differential validity) much earlier in the sequential selection process (Helme, 1965). The non-cognitive tests in the ARI battery were primarily self description tests with empirically developed keys using performance in several different areas as the criteria — compared to the factor based personality tests in the Project A battery that were not included among the first 9 tests selected from the Project A battery. The interest tests (AVOICE) of the Project A battery faired fairly well but none were selected as early as the non-cognitive tests in the earlier ARI battery, while none of the "job orientation composites" (JOB) nor "temperament and biodata composites" of the Project A battery were in the first 10 selected (see *Personnel Psychology*, 1990).

We are definitely not saying that the existing cognitive test battery would not have its classification benefits increased by the addition of appropriate non-cognitive tests or measures of job knowledge or vocational aptitude focused on vocational domains differentially relevant to Army jobs (e.g., electrical, mechanical, automotive, engineering technology). We are saying that the kind of non-cognitive variables included in the earlier batteries (e.g., Helme) are better predictors of performance criteria than are tests based on personality theory.

Summary of Objections

The authors of the HumRRO Report present DAT in such a way that it is easy to challenge whether DAT supports the various benefits alleged to be provided by the TTCS. The HumRRO report appears to adopt the strategy of the early "*g*" theorists who proclaimed that there is "little more than '*g*' " in valid predictors, and suggest Horst and Brogden are supporters of the specific factor theory. According to the "*g*" theorists, benefits from the use of multiple test

batteries in classification and assignment rely on the successful matching of specific factors to job families. Since few researchers give much credence to such a specific factor theory, including Horst and Brogden, and more recently, others who give credence to DAT, the accusation that a theory or process relies on the use of specific factors is a harsh indictment that should be proved whenever made.

The HumRRO authors argue for an evaluation model that is appropriate for determining the efficiency of selection, but not for personnel classification. Using this argument they claim that samples used to compute regression weights for a composite must each be at least 2,000 to avoid validity shrinkage in the cross sample to an extent that would remove all benefits provided by "best" weights for test scores, reducing cross validity coefficients to the level provided by "*g*". They rely on the examination of standard errors, instead of on cross sample results, to reach this conclusion.

The HumRRO authors make frequent reference to the impact of GMA (i.e., "*g*") in their arguments against expanding the number of composites beyond 9. Some of their arguments are similar to Air Force "*g*" theorists who appear to be proposing the use of a single family with a single composite that measures "*g*" (Ree & Earles, 1994). The AF still uses 4 job families with corresponding composites. We assume the number of composites was originally reduced to 4 to correspond to the number of non-trivial group factors that results when the correlations among the operational sub-tests are factored using a PC solution and then rotated to simple structure.

Surprisingly, the HumRRO authors do not deal with the obvious problems that come with the alternative to making use of empirical small sample data for computing R and r. The alternative is to rely on judgment to determine which of the 9 operational composites best predicts performance in the MOS represented by small families. We have no qualms with using

judgment to reduce 250 families to 150 families by aggregating the 100 families with the smallest *N*s with the 150 families that were formed by using empirical data to minimize MPP in the clustering process. There is almost always a job family which is clearly very similar to the small-*N* MOS being considered for placement. In contrast, we don't feel that one can justify using judgment to assign each of the 250 MOS to one of the 9 operational job families, and to then use the optimal assignment process to assign recruits to one of the 9 operational job families and to one of the 250 MOS. The latter step reflects the judgmental relationship between the MOS and the nine job families.

Conclusions and Recommendations

The HumRRO authors display a serious misunderstanding of the structure and purpose of the TTCS, and misinterpret DAT at several points. In addition, they appear to be overly influenced by the "little more than '*g*' " dogma of the "*g*" theorists as the basis of their evaluation of the second tier battery (Ree & Earles, 1991). They appear to be making an effort to resell the "*g*" theorist position that only measures of general ability are stable enough to form the core of an operational test battery. Their misunderstanding of the TTCS and DAT, combined with their devotion to this dogma, makes many, if not most, of their recommendations pertaining to TTCS either inappropriate or wrong. However, it appears that they are not completely converted to "*g*" theory dogma, since they seem to agree with us that LSE weights applied to each test in the battery are appropriate for use in the first tier battery.

The HumRRO authors made a valuable contribution in their earlier report by checking the estimation of the second tier composite regression weights. For years we have been advocating that the regression weights for the nine standardized composites, constrained to be

positive, be independently checked, and recently ARI contracted to have this check made by HumRRO. Future researchers should seriously consider using their multi-sample cross-validation design to compute MPP for each job family and condition. However, the HumRRO research design should not be used to compute standard errors of MPPs because of the overlap of their analyses and evaluation samples across each pair of Sample C sub-samples.

The HumRRO authors express a number of opinions that could be tested empirically, and simulation experiments would be appropriate for examining some of these opinions. The first set of such opinions relate to the relationship of the number of composites in a configuration (m) to the magnitude of MPP. The HumRRO authors failed to consider that the MPP obtained for configurations with $m = 9$ and $m = 17$ are considerably less than would be provided by optimally clustering MOS into job families using the Horst algorithm. Families formed by judgment affected by administrative and political considerations provide lower values of MPP than those formed using Horst's algorithm in forming job families. We believe this is still true when job families are formed in the analysis sample – i.e., in the same sample in which the regression weights are computed for the AA composites. While we agree with the HumRRO authors that some inflation in MPP results from using job families optimally clustered to maximize differential validity in the total sample (rather than in Sample A), we will be surprised if this inflation is great enough to seriously affect the gains in MPP provided by using more than 9 job families for classification and assignment.

The change in MPP as m increases should have been computed using optimally clustered configurations for $m = 9$ and for each other value of m between 9 and 150 for which MPP is obtained. The HumRRO authors' interpretation of the meaning of the changes in differences in gains in MPP between different values of m would have made more sense if they had obtained

such empirical results. Zeidner and colleagues have examined such a continuum (Zeidner, et al., August 2000, p. 51) and have found that the increase in MPP gradually decreases but remains positive with the increase of m until somewhere around $m = 80$ where the increase of MPP becomes negative and MPP starts to very gradually decrease. This should be confirmed by using a design that includes forming families in the same sub-sample (A or B) that is used to compute regression weights.

The second set of opinions that should be empirically checked relates to the HumRRO authors' statements regarding the effect that operationally applied constraints would have on the benefits of values of m greater than 9. The authors appear convinced that the operational application of quality constraints regarding key MOS would eliminate the gain in MPP provided by using non-standardized composites and/or a higher value of m. This can be readily checked by a simulation experiment in which the optimal assignments are constrained to meet quality constraints in key MOS, in addition to meeting MOS quotas.

Early in the report the HumRRO authors appear to base their conclusions regarding the probable benefits of using more than 9 job families and associated composites to make initial classification and assignment of recruits largely on the basis of the larger standard error of R obtained in the smaller MOS, supported by comparatively vague references to "bouncing betas", high values of r, and small sample size (as small as 200 in several families when $m = 150$) in our data set for a few of the 150 families.

The HumRRO authors' use of analysis of variance tests of the significance of differential validity coefficients (R) to determine classification efficiency does not comprise a defensible substitute for comparison of MPP across conditions of interest. For example, the sampling distribution of R when families are based on judgment is not readily comparable to when the

families are obtained by maximizing R. Also, classification efficiency as determined by MPP is a function of the average magnitude of composite inter-correlation coefficients (r), and a function of m (as described by Brogden) as well as R.

We recommend that ARI conduct simulation experiments in which optimal assignments are constrained to meet both MOS quotas and quality constraints for first tier configurations of m = 9, 25, 40, 66, 75, and 150 in both the standardized and non-standardized versions of PPs. The clustering of MOS into families should be accomplished separately in Sample A and Sample B. The use of either our triple cross validation design or HumRRO's modified design could be used for computing MPP, but not for computing the standard errors of MPPs. Either way optimal constrained assignments would be made in Sample C sub-samples. If results are sufficiently promising, a second experiment – which uses an algorithm equivalent to the doubly constrained linear program – should be conducted to demonstrate an operationally practical algorithm for making recommended doubly constrained optimal assignments for use by counselors.

We also recommend that the classification efficiency of the non-cognitive measures in the Helme experimental test battery (Helme, 1965) be compared to that provided by the non-cognitive tests in the Project A battery and formed (with the additional tests obtained by implementing the most promising of the non-cognitive test concepts proposed by Helme and other in-house ARI scientists) into a battery of experimental tests for a research effort to sequentially select the best tests for improving classification efficiency. Such research could greatly improve the classification efficiency of the ASVAB.

References

Brogden, H. (1955). Least squares estimates and optimal classification. *Psychometrica*,20, 244-252.

Brogden, H. (1959). Efficiency of classification as a function of number of jobs percent rejected, and the intercorrelations of job performance estimates. *Educational and Psychological Measurement,* 19, 181-190.

Diaz, T., Ingerick, M., & Lightfoot, M.A. (August 2004). *Replication of Zeidner, Johnson, and colleaques' method for estimating Army Aptitude Areas (AA) composites* (Study Report 2004-04). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Diaz, T., Ingerick, M., & Lightfoot, M.A. (December 2004). *Evaluation of alternative aptitude area (AA) composites and job families for Army classification* (Study Report 2005-01). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Helme, W.H. (June, 1965). Evaluation of differential classification for the ACB. *BESRL Technical Research Note 155.* June 1965 (AD 621-698). Washington, DC: BESRL.

Horst, P. (1954). A technique for the development of differential prediction battery. *Psychological Monographs, 68* (69, Whole No, 380) 1-22.

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

Johnson, C.D., Zeidner, J., & Leaman, J.A. (1992). *Improving classification efficiency by restructuring Army job families* (TR-947-92). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, C. & Zeidner, J. (July, 1995). *Differential Assignment Theory Sourcebook* (Research Note 95-43). Alexandria, VA: United States Army Research Institute for the Behavioral and Social Sciences.

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

Personnel Psychology (1990, Summer). Project A: The U.S. Army Selection and Classification Project (Special Issue), 43, 2.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44,* 321-332.

Ree, M. J. & Earles, J. A. (1994). The ubiquitous predictiveness of *g*. In M.G. Rumsey, C.B. Walker, & J.H.Harris (Eds.) *Personnel selection and classification*. Hillsdale, NJ: Erlbaum Associates.

Schmidt, F.L., Hunter, J.E., & Larson, M. (1988). General cognitive ability vs. general and specific aptitudes in the prediction of training performance. Some preliminary findings. (Contract No. Delivery Order 0053). San Diego , CA: U.S. Navy Personnel Research and Development Center.

Zeidner, J., Johnson, C.D., & Scholarios, D.M. (1997). Evaluating military selection and classification systems in the multiple job context. *Military Psychology*, *9*, 169-186.

Zeidner, J., Johnson, C.D., Vladimirsky, Y., & Weldon, S. (2000). *Specifications for an operational two-tiered classification system for the Army, Volume 1* (TR-1108-VOL-1). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Zeidner, J., Johnson, C.D., Vladimirsky, Y., & Weldon, S. (2003a). *Determining composite validity coefficients for Army jobs and job families* (SN-2003-02). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Zeidner, J., Johnson, C.D., Vladimirsky, Y., & Weldon, S. (2003b). *Determining mean predicted performance for Army job families* (SN-2003-03). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.