# Measuring Job Competency (1988)

Pages
37

Size
5 x 9

ISBN
0309320070

Green, Jr., Bert F. and Wigdor, Alexandra K., Editors; Committee on the Performance of Military Personnel; Commission on Behavioral and Social Sciences and Education; National Research Council

# Measuring Job Competency

Bert F. Green, Jr., and Alexandra K. Wigdor, *editors*

Committee on the Performance of Military Personnel
Commission on Behavioral and Social Sciences and Education
National Research Council



NATIONAL ACADEMY PRESS
Washington, D.C. 1988

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Frank Press is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Samuel O. Thier is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

The work of the Committee on the Performance of Military Personnel is sponsored by the Office of the Assistant Secretary of Defense (Force Management and Personnel) and funded under Office of Naval Research Contract N00014-83-C-0448.

Available from:
Committee on the Performance of Military Personnel
National Research Council
2101 Constitution Avenue N.W.
Washington, D.C. 20418

Printed in the United States of America

## COMMITTEE ON THE PERFORMANCE OF MILITARY PERSONNEL

# Preface

The Committee on the Performance of Military Personnel was formed by the National Research Council, principal operating agency of the National Academy of Sciences and the National Academy of Engineering, to provide scientific oversight to the Joint-Service Job Performance Measurement/Enlistment Standards Project. The Joint-Service Project is a large-scale demonstration project to measure the performance of first-term enlisted personnel in the Army, Air Force, Navy, and Marine Corps, and to link enlistment standards to performance on the job. It includes the development and administration of an array of performance measures for approximately 28 military jobs and the development of techniques for making the resulting performance information useful to the policy makers who set standards. Although the research is being carried out at the Service level, the Office of the Assistant Secretary of Defense (Force Management and Personnel) provides overall direction and coordination of the project.

At the request of its sponsors in the Department of Defense, the committee has since its inception given attention to the potential usefulness of the research for personnel decisions and manpower management. Some Service personnel have tended to a narrow view of the Joint-Service Project as simply a validation of current entrance tests (the Armed Services Vocational Aptitude Battery or ASVAB) using job performance as the criterion rather than training success. The Department of Defense recognizes the

importance of this aspect of the studies, but anticipates that the project can also provide a scientific basis for validating entrance standards, for allocating personnel to various military jobs, and for justifying the Services' quality requirements to Congress.

The committee became convinced early in its existence that these larger project goals make the interpretation of scores on the job performance tests of fundamental importance. It has recommended strongly that the Service research teams attempt to develop and score the performance measures so as to permit a competency interpretation of performance scores. The rationale for this recommendation is described in an earlier report (Wigdor and Green, 1986). Following the publication of that report, selected members of the committee and staff participated in a series of meetings with members of the Job Performance Measurement Working Group (which, under DOD chairmanship, coordinates the Joint-Service Project) to discuss the meaning of assessing competency and to devise ways of establishing such interpretations and using the results. (A list of participants in these joint discussions follows this preface.)

In writing about this issue, we have chosen the word *competency* with intent. The word *competence* is frequently used to indicate a categorical characterization that contrasts with incompetence. We mean competency to encompass the entire range of performance, from not at all competent to extremely competent, with many levels in between. In fact, this perception of competency, and its contrast with a dichotomous concept, is one of the positive outgrowths of our joint discussions.

This report grew out of and supplements our lively deliberations with members of the Job Performance Measurement Working Group. Although the discussants reached consensus on many aspects of the question of measuring job competency, the committee is more convinced of the importance of attempting to assess competency than some of the military participants are convinced that it is feasible for them to do. Consequently, this report, while hereby acknowledging its debt to the various Service contributors, is presented as the unanimously endorsed position of the committee and is earnestly presented for the serious consideration of the Job Performance Measurement Working Group.

Bert F. Green, Jr., Chair
Committee on the Performance of Military Personnel

# PARTICIPANTS

## Meetings on the Assessment of Competency
## October 24-25, 1986; March 13-14, 1987

**Committee on the Performance of Military Personnel:**

Bert F. Green, Jr., Department of Psychology, Johns Hopkins University
Robert L. Linn, School of Education, University of Colorado
Richard J. Shavelson, Graduate School of Education, University of California, Santa Barbara
Alexandra K. Wigdor, National Research Council

**Job Performance Measurement Working Group:**

Jane M. Arabian, Army Research Institute for the Behavioral and Social Sciences
Robert L. Frey, Headquarters, Coast Guard
Lt. Col. Dickie A. Harris, Directorate for Accession Policy, Department of Defense
Jerry W. Hedge, Air Force Human Resources Laboratory
Gerald J. Laabs, Navy Personnel Research and Development Center
Jerry Lehnus, Defense Manpower Data Center, Department of Defense
Milton H. Maier (Marine Corps), Center for Naval Analyses
Paul W. Mayberry (Marine Corps), Center for Naval Analyses

# Measuring Job Competency

## THE RECOMMENDATION TO MEASURE COMPETENCY

The Job Performance Measurement/Enlistment Standards Project of the Armed Services was established to examine the feasibility of measuring job performance and to link enlistment standards to job performance. The Committee on the Performance of Military Personnel, which was established to provide technical oversight to the project, expects the project to demonstrate several methods of measuring job performance adequately. The process of linking entrance standards to job performance is a more complex task requiring nontraditional methods and an expanded sense of policy perspectives.

The committee feels strongly that if the Joint-Service Project is to effectively communicate information about the performance of enlisted personnel and the implications of changing standards—either internally to military policy makers or to Congress—then the scoring scale of the job performance tests needs to be given some sort of absolute meaning. Scores should, in other words, communicate some sense of how well a person can do the job or, perhaps, how much of the job a person can do well. In contrast, scores currently say something about an examinee's relative standing with reference to all other examinees, which is useful for ranking applicants but is not very informative about how a person at any particular score level will perform a given job. Measures of job competency would need to be referenced to some external

scale of job requirements, not to the performance of other job incumbents.

The term *competency* as used here denotes a way of interpreting scores on a performance scale. It follows that there are degrees of competency. Unfortunately, the term has sometimes been used to signify a simple dichotomy, separating the competent from the incompetent.

That is not our meaning, nor our intent. As we shall argue, a performance dichotomy is neither implied nor necessary. In selection systems, minimum standards or cutoffs are placed on entrance tests, not on performance measures—on the input, not the output. Setting a particular input standard will result in a consequent output distribution of job performance scores, some low, some intermediate, some high. Policy makers must decide if the resulting distribution of performance scores is acceptable. They would be better able to make informed judgments about what is acceptable and what is unacceptable if performance scores could be interpreted in terms of what the job incumbent who scores at each level is able to do.

## Performance-Based Selection Standards

To clarify this point, we sketch a very simple model for setting entrance standards. This basic analysis leaves aside many considerations and is provided only to illustrate the relationship between selection test cutoffs and performance scores.

The general problem in all entry-level jobs is how to cope with a distribution of proficiency. Inevitably, some incumbents will perform poorly. Technical training schools cannot be expected to turn out only experts. A more realistic expectation is that job incumbents will develop and improve on the job. There is always a flow of personnel through a job. As some incumbents become experts, others are being promoted or released, and still others are just entering the job. There will always be some novices, some apprentice-level job incumbents, and some experts (given sufficiently stringent enlistment standards). For manpower management, it would be very desirable to establish an expected or realistically acceptable distribution of proficiency in a job cadre.

Figure 1 shows predictor composite scores and performance scores that are related in the usual psychometric fashion, assuming

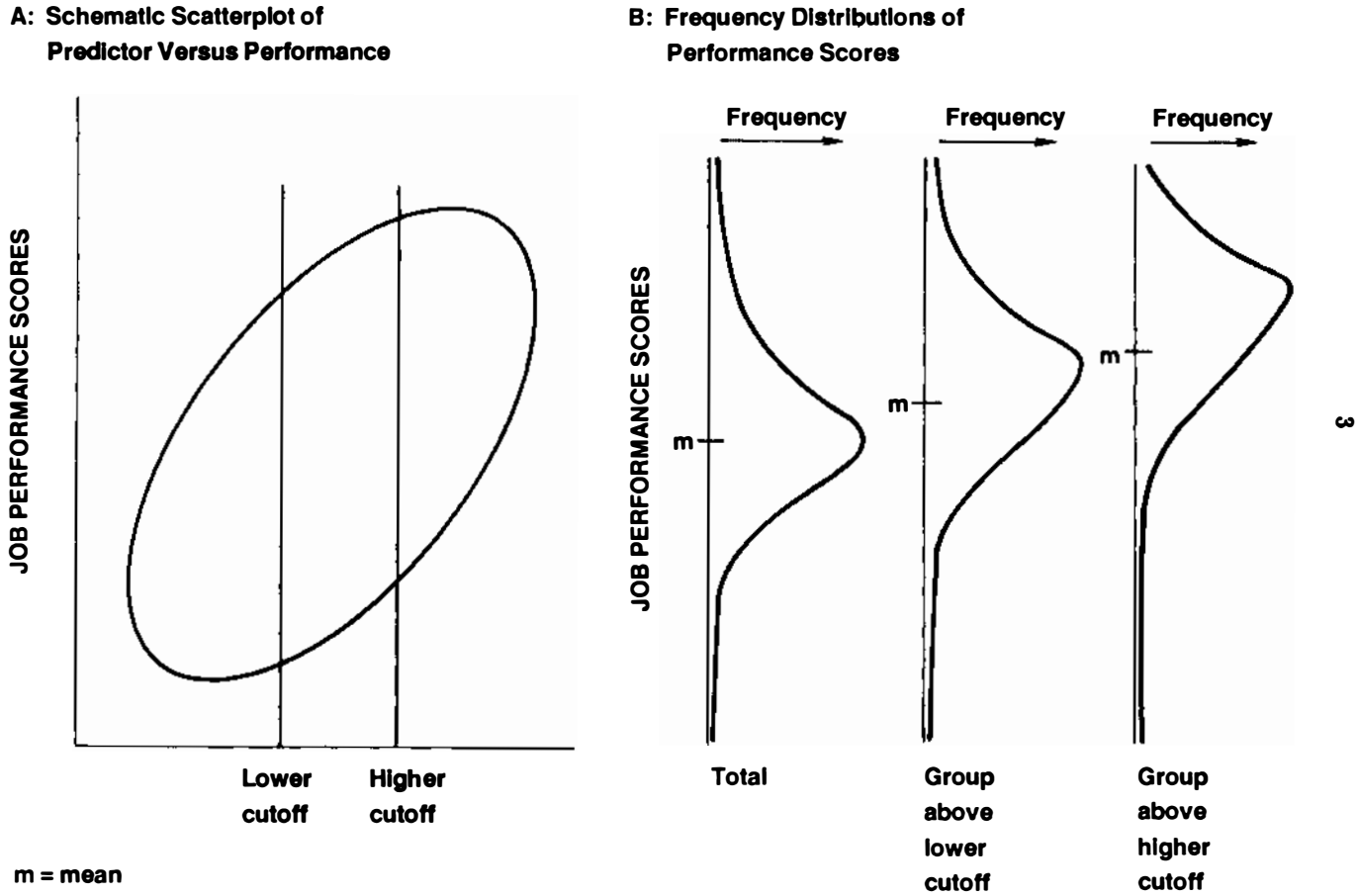


FIGURE 1 Scores on predictor composite with resultant performance distributions.

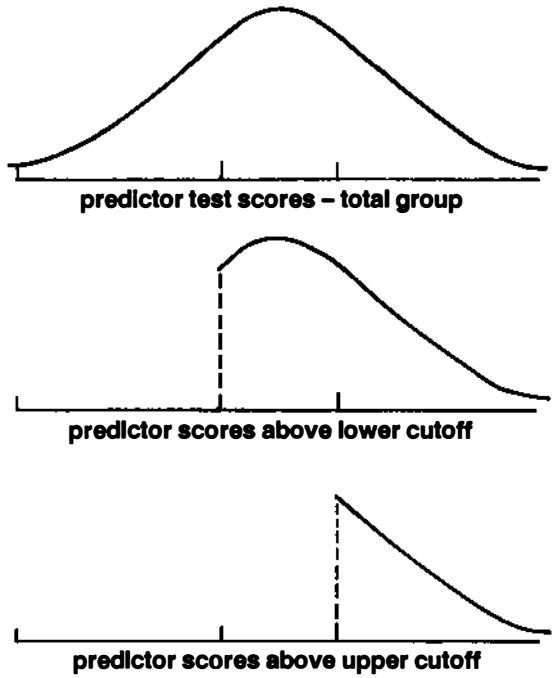


FIGURE 2 Distribution of predictor composite scores.

a moderate validity correlation and roughly normal score distributions. The population is considered to be those who actively seek the job in question. For purposes of discussion, we assume the availability of performance scores for persons who will not be selected and therefore will have no chance to actually perform. Each person is in principle represented in the diagram by a point relating their predictor score with their performance score. The set of points forms a swarm roughly elliptical in shape, as indicated by the ellipse in Figure 1A. Two different standards or cutoffs are depicted on the predictor measure. As Figure 2 shows, each standard cuts off a group of scores and leads to a distribution of the performance scores that exceed the cutoff. Note that the distribution of performance scores arising from the more stringent cutoff on the predictor has a higher mean, a smaller spread, and greater negative skew (Figure 1B).

Two major points are clear from this schematic view of the selection process. First, setting a cutoff on the predictor composite

does not entail setting a corresponding cutoff on the performance measure. (Setting a minimum acceptable performance score would certainly be possible, but it would be a separate step.) The converse is also true: setting a minimum acceptable performance does not imply a corresponding cutoff on the predictor.

Second, evaluating the result of a particular predictor cutoff requires evaluating the resulting distribution of performance scores. Whether a given cutoff is acceptable depends on whether the corresponding performance distribution is acceptable, as well as on the additional considerations of cost, manpower needs, etc. To decide whether to accept a certain performance distribution, both policy makers and modelers need some way of interpreting performance score distributions—the committee argues for an absolute referent through a competency-based scale. Furthermore, the entire distribution is at issue, not just some acceptable minimum performance level.*

## Interpreting Performance Scores

The interpretation of a performance test score refers to the inferences about job performance that can legitimately be drawn from criterion test performance. To the extent that a criterion measure is representative of the work required on the job, some

*A standard way of explaining validity correlations is by way of expectancy charts. Such charts are based on a dichotomous view of performance: success versus failure. The chart shows the proportion of candidates at each predictor score level who may be expected to succeed or pass. Sometimes *succeed* has an objective meaning, but in the current arena of performance measures, it does not. Rather than identifying some minimally acceptable performance level, we suggest that the entire performance distribution should be evaluated. In this case, expectancy charts are oversimplifications.

Apart from such diagrams, traditional studies of the validity of predictor tests generally ignore the issue of minimum standards or cutoffs on the predictor scores. It is more or less implicitly assumed that the highest scorers are selected. If people arrive in batches, as in the yearly batch of college entrance applicants, then cutoffs may be ignored. However, when persons are applying daily, as in the military, minimum standards are necessary. In fact, of course, minimum standards are useful and are often used in batch processing too, because the selection process always involves more elements than potential performance. In college admissions, for example, a student whose parents are alumni, or who plays football well, or who comes from an underrepresented part of the country might get extra consideration, but only if the student is predicted to achieve at least a passing grade-point average.

kind of inference is warranted from the test to the job domain. Thus, the process of investing a performance score with meaning begins with a careful study of the job and involves selecting tasks for testing that adequately represent the entire domain of job requirements.

The most straightforward (if also simplistic) procedure would be to start with an inventory of job tasks and to sample randomly from that list to form the test. Given a test of sufficient length, a person who could do 70 percent of the tasks on the test would be expected to be able to do about 70 percent of the tasks in the job. Sophistication can be built into the sampling design by clustering tasks and weighting those clusters to mirror what is adjudged to be the "essential" job. For example, tasks might be organized into functional groupings, representing different dimensions of the total job, or they might be organized according to the types of behavior they require. Tasks might also differ in importance, difficulty, or frequency. If these factors are considered important to the definition of the job, and if they can be made explicit, they can be included in the sampling and estimation procedure.

Whether a random or purposive sampling scheme is adopted, it is clear that the initial definition of the job domain is the foundation of any later interpretation of performance test scores. Without some demonstrable claim to representativeness, performance test scores (and criterion measures in general) have little or no meaning.

The kind of inferences that can be drawn from performance test scores are also affected by the dictates of psychometric techniques. For some measurement purposes, e.g., validation of predictors, the central aim is to demonstrate individual differences in performance. Tasks are selected from the middle range of difficulty—neither so easy that everyone performs them correctly, nor so difficult that no one does so—in order to produce a distribution of scores. As a result, the representativeness of the instrument may be qualified by the psychometric goal of spreading performance across a broad continuum. The resulting test score is a norm-referenced score, the norm being the population of test takers. Norm-referenced test scores have only relative meaning. For example, a person with an ASVAB standard score of 50 on the word knowledge test has a working vocabulary about as extensive as the average applicant, but apart from this relative statement,

the score indicates nothing about the extent or adequacy of his or her vocabulary.

It is appropriate for predictor tests to be scored to show relative standing in the tested population. The committee has argued, however, that criterion scores that allow only a normative interpretation, while useful for examining the validity of a predictor composite, have a limited use beyond that. For example, the validation of a selection standard, i.e., a minimum cutoff on the predictor composite, requires an evaluation of the resulting expected performance along the distribution of selected applicants. Knowing that a higher score implies better performance is not terribly informative. What is needed is a sense of how good the performance is at points along the scale. This implies having an externally defined scale of performance with scores referenced not to the relative performance of others but to levels of job mastery.

## Domain-Referenced Testing

After much discussion, both substantive and semantic, the participants in the meetings on competency measurement agreed that domain-referenced testing is the most suitable vehicle for a competency-oriented approach to job performance measurement. The essential feature of domain-referenced testing is that its interpretive framework is not a population of test takers, but rather a content area, e.g., the tasks in the job of a jet engine mechanic. With domain-referenced testing, the interpretation of test performance has to do with what the examinee knows or can do, not how he or she compares with other examinees (e.g., Brennan, 1981; Shavelson and Webb, 1981). If the test adequately represents the domain of interest, it can be scored to indicate how much of the content domain has been mastered. For example, a student scoring 70 on a domain-referenced final examination in intermediate French vocabulary can be assumed to know about 70 percent of the words in the specified domain of French vocabulary, as defined by the content of the course (e.g., Lennon, 1956).

The term *domain-referenced testing* was chosen instead of the more commonly used *criterion-referenced testing* to make an important distinction. Although both terms imply a content-referenced interpretation of test performance, criterion-referenced testing has become closely associated with minimum competency testing programs in recent years. In numerous states, high school

students are required to demonstrate minimum levels of competence in language skills, mathematics, and possibly other areas of local or state interest as a prerequisite to graduation. The purpose of this kind of testing is to determine if the student has met the required minimum level of performance. Rather than specifying any kind of performance level on a domain, evaluators define a minimum level *on the test scale*. Once the criterion level has been defined, there is little interest in differentiating among degrees of success or degrees of failure.

Thus, for purposes of minimum competency testing programs, a good criterion-referenced test is designed to differentiate well at the critical criterion level, and not so well elsewhere on the scale. That is, all the items are of about the same difficulty and are chosen to represent the minimum level of competency as defined by the educational specialists. By contrast, domain-referenced tests need not involve minimum performance requirements, and the meaning of the scores must be understood throughout their range.

## Advantages of a Domain-Referenced Scale

The major advantage of supplying externally referenced meaning to the score distribution is sheer interpretability. The Joint-Service Project was occasioned by a scoring problem that inflated scores, leading, among other things, to the erroneous induction of 250,000 enlistees who did not actually meet the mental ability standard. Whereas it was thought that 5 percent of enlisted accessions in the period 1976 to 1980 were in Armed Forces Qualification Test (AFQT) Category IV, the lowest category of eligibility, later corrections indicated that the figure was more like 30 percent (Maier and Truss, 1983). Personnel managers and military and congressional policy makers were understandably concerned to find that the same scores meant different things in 1975 and 1976. Some were doubly concerned to realize that the existing technology was not of much assistance in understanding scores in a more substantive sense (beyond form-to-form equivalence). The tools were simply not at hand to describe the kinds of performance deficits that might be expected across the distribution of new accessions. Therefore, it was difficult to estimate the significance of the problem.

The ultimate goal of the Joint-Service Project is to link enlistment standards to on-the-job performance. This goal can be interpreted more or less expansively, with commensurate benefits. If predictor scores can be correlated with scores on the performance measures—and preliminary data analysis indicates that a reasonable correlational relationship exists—then the discussion will have been advanced on all sides. But the potential payoff for military manpower, personnel, and force management systems will be far greater if the research goes beyond correlational analysis. At least four purposes can be distinguished for examining the linkage between enlistment standards and job performance, and, in the committee's judgment, three of the four could be enhanced if the research were based on a competency approach. Among the purposes that have been discussed are:

1. Demonstrating that selection instruments have validity for predicting job performance;
2. Providing empirical information for setting enlistment standards;
3. Providing performance information for making allocation decisions;
4. Providing performance-based estimates of force quality requirements.

Each of these represents an important step in strengthening the scientific basis of military manpower and personnel policy. Together, they promise significant improvement.

## Validity

The first purpose, establishing predictive validity, is the easiest to accomplish. It does not require a competency approach for its success. At present, the predictive validity of the ASVAB for training school success is well documented (U.S. Department of Defense, 1985:iii), but validity for actual performance is only assumed. The Joint-Service Project will examine predictive validity by correlating entrance test scores with job performance scores. There is every reason to expect adequate validity, but the evidence might reveal a different pattern of validities of various ASVAB subtests for actual performance.

## Entrance Standards

Although predictive validity shows the efficacy of selection tests in predicting job performance, it does not speak directly to the question of entrance standards. In order to enable those who set entrance standards to take expected job performance into account in a systematic way, the Joint-Service Project will need to establish the performance effects of changing entrance standards (and therewith the ability mix of job incumbents). It is here that competency scaling becomes significant. If the scaled performance scores signify the levels of job proficiency to be expected, then the relation of performance scores to selection test data will give military policy makers far better information than they now have for setting standards for entry into each occupational specialty. Using that information will still involve difficult decisions about acceptable distributions of proficiency, together with other considerations; competency measurement does not solve the problem of standard setting, but it does provide a sound base.

## Allocation

Once a recruit has qualified for enlistment into a Service, the jobs of that Service are in competition with each other for the enlistee. Job allocation systems attempt to decide the relative benefit to the Service of the recruit's various job options. As far as the committee has been able to ascertain, predicted performance is not a significant factor in the current allocation algorithms; the systems are driven more by management objectives, particularly fill rates, than by considerations of job performance.

If performance measurement is to be useful in allocation systems, performance scales will need to be translated to a common metric so that competing jobs can be compared. A competency analysis seems to the committee a particularly fruitful way to approach the problem of comparing jobs, since the competency designations developed for each job's performance measures could be correlated with the predictor tests given at entrance and could guide allocation. For example: if enlisted personnel in Job X who scored in the 50th percentile on the relevant ASVAB technical composite consistently achieve expert status by the end of the first term, one would want the allocation system to avoid waste by

not assigning people to Job X if their score on the technical composite is very much above the 50th percentile. Likewise, if similar enlisted personnel assigned to Job Y tend to hover at the apprentice level of mastery at the end of the first term, one would want the system to tend to avoid sending applicants to Job Y whose composite score was much below the 50th percentile, despite fill rate deficits. Since the military allocation systems are all computerized, it is possible to accommodate such complex decision situations.

Rather than making the job performance measurement research directly applicable to military allocation systems, an interesting "end-run" around the competency issue might be possible if jobs can be put on some other single metric. The allocation system could be set to pick jobs for which the applicant has the highest scores on this common scale. Each Service has been exploring ways to produce a common scale that will allow comparisons across jobs. The Air Force learning difficulty research offers the possibility of comparing occupations on the basis of the time it takes to learn primary tasks, which is taken as a measure of the difficulty of a job. The Army has studied a utility scale using both officer and noncommissioned officer judgments in an attempt to assess the usefulness to the Army of a person with a certain performance test score. The utility scale is intended to permit comparisons across jobs. The Marine Corps plans to gather judgments of value, on some sort of scale, as a precursor to applying a manpower model, e.g., the RAND model, for setting cutoffs on the aptitude composite score to optimize payoff from the personnel system.

Each of these scales has merit, but none provides a direct index of proficiency. These scales cannot be used to allocate applicants in terms of how well they can be expected to do their military jobs. Learning difficulty addresses a different problem, the design of training courses. Utility and value scales mix the concept of proficiency with value judgments in undetermined ways. Value and utility are certainly of concern, but these concepts should follow upon competency measurement, being judgments of the expected proficiency, rather than being integral to the measurement. Competency is a first step; the utility and value of various levels of proficiency should be determined as a subsequent step. The currently conceived utility scales bypass the concept of performance, and might even be said to disguise competency, which is central to

the military mission. The committee feels that competency is an essential component of any method of balancing the needs of the competing occupational specialties.

## Quality Needs

From the point of view of the Office of the Secretary of Defense (OSD), one of the most important contributions of the Joint-Service Project will be the increased precision with which the Services can estimate their quality needs. As part of the Omnibus Defense Authorization Act of 1985, the Senate Armed Services Committee required the Department of Defense to review military enlisted manpower quality requirements for the next five years. In order to make these projections, the Services had to rely on two indirect indicators of quality: high school education status and scores on the Armed Forces Qualification Test. Although high school graduates are far more likely to complete the first term of enlistment than nongraduates, and AFQT scores are positively correlated with scores in technical training, OSD looks to the job performance measurement research as a source of more direct and therefore more credible evidence of the Services' quality needs. The Joint-Service Project will relate entrance quality directly to job proficiency, a critical component of force readiness.

The ultimate goal in projecting quality requirements is to balance performance gained against the costs of recruiting, training, and retaining personnel. A competency-based assessment of performance would be of obvious value in understanding—and allowing Congress to understand—the effects of increasing or decreasing the money budgeted for recruiting, training, benefits, or other personnel costs. It would add credibility to what is necessarily a very complicated judgment.

One of the more complicated problems being explored by the research teams is how to incorporate performance needs and manpower costs in some sort of trade-off model that would permit the evaluation of the relative costs and benefits of differing force quality levels. Ideally, such a model would be responsive to labor market conditions, the recruiting climate, budget realities, and changes in the nature or difficulty of military jobs. It would help policy makers by locating the optimal quality mix to minimize the total cost of recruiting, training, and maintaining the force. Early experiments with manpower management models used a

norm-referenced performance factor and simply chose arbitrarily a minimum standard to define the performance objective. A competency-referenced performance factor with known proficiency distributions would have the great advantage of preventing costs from driving the model to the point at which individual proficiency suffered. If performance expectations could be determined for all jobs and if proficiency distributions could be compared on a common scale, DOD would be better able to justify its projections of force quality requirements.

## Ancillary Uses of the Job Performance Measures

Service representatives voiced a concern with how the research results of the Joint-Service Project will be used. The initial characterization of the Joint-Service Project was as a research effort. First, the Services were to see if it is feasible to develop good measures of job performance; assuming the success of that activity, they would study a variety of methods of linking proficiency scores to enlistment standards to see how the performance data could be operationally useful. The impetus for the research and its conceptual focus was on selection and classification issues.

Since then, the promise of the new performance measures has spurred interest in other possible applications, e.g., to make promotion decisions, to evaluate training effectiveness, or to assess the combat readiness of units. To the extent that such discussions focus on the potential usefulness of the technology per se, the Joint-Service Project should be a fertile source of information on new methods and new assessment tools. However, the current instruments have not been designed for the ancillary applications, and we fear that expanding the use of these very job performance measures beyond the original intention of evaluating alternative enlistment standards could pose serious threats to their measurement validity.

One type of problem is test fairness. In the current research environment, test-takers can be promised anonymity; the test outcome will not be part of their individual personnel record. Judging from our extensive site visits, the test-takers do their best under these circumstances, but there is probably little motivation for prospective test-takers to find out what will be on the test, nor for those tested to pass on details. Thus, the test appears to provide a valid indication of what incumbents can do. However, if

the performance data were used for making decisions that affect the welfare of individuals—the test subjects, their supervisors, unit commanders, or teachers of technical training, for example—there would be a strong inclination for people to try to protect their interests. Coaching the test-takers would be inevitable. There is no way to avoid having the content of the current tests known, at least in general terms, to prospective test-takers. Because the proficiency tests are only limited samples of the domain of job requirements, any resulting improvement on the tested sample could not be assumed to translate to an improvement in total job performance. The validity of generalizing from the test scores to overall job proficiency would be seriously threatened.

Attempting to develop job performance measures that would serve administrative functions over and above the Joint-Service Project goals could also raise problems of test content. For measuring individual job proficiency, test content should represent the domain of job requirements; for evaluating training effectiveness, test content would ordinarily focus on the training objectives; estimation of combat readiness would require more than measures of individual proficiency. Although the measurement technologies that were developed in this project could be used to construct tests for different applications, the present instruments would probably not be suitable.

There was general agreement that (1) attempts to expand the uses and interpretations of the Joint-Service Project performance measures beyond the intended applications in personnel selection and classification require a thorough evaluation of the appropriateness of the measures to each additional proposed application and (2) that it would be ill-advised to threaten the validity of the Joint-Service Project performance measures either by using them in ways that could affect individuals' careers or by attempting to make them all-purpose measures.

## OPERATIONALIZING THE COMPETENCY IDEA

Having explored the rationale and potential benefits of a competency approach to job performance measurement, participants in the meetings on competency assessment took up the practical question of how to develop measures that permit interpretation of performance scores as representing degrees of job competency or job mastery. For this specific application, the fundamental need

is for the measures to be representative of job requirements. The problems include representing the job domain, scoring the test, and providing interpretive anchors for the resulting scale. The approach to each problem was guided by the eventual goal of accurately expressing the individual's level of proficiency, rather than maximally discriminating among individuals.

## Representing the Job Domain

The first step in defining a scale of competence is specifying the domain of the performance being measured. A competency scale is defined by assuming the existence of a finite, specifiable measurement domain, in this case a job or occupational specialty. Job elements must be defined as a preliminary to test construction. The process of job specification requires decisions about the boundaries of a job and the most appropriate units of analysis. In the Joint-Service Project, the Services have defined each job in terms of its component tasks and have used job tasks as the appropriate units.

Once the domain has been specified, a sample of the tasks can be chosen as a basis for creating the performance measure. Because the tasks in a job can often be clustered into types of tasks, there would be merit in stratifying the tasks in accordance with those clusters and adapting the sampling procedure to match the job organization by sampling each stratum separately, in a frequency perhaps proportional to the sizes of the various strata.

Other factors than the organization of the job can reasonably be used in defining the strata or in establishing sampling weights. In particular, tasks also differ in importance, difficulty, and frequency. If these factors can be made explicit, they can be included in the sampling procedure.

If, as we typically assume, jobs are multidimensional, the problems of job specification increase. The difficulty factor provides an illustration. The concept of difficulty in the context of selecting test content implies a rank order of skills and knowledge; that is, people who can perform the more difficult tasks can also perform the easier ones. In a unidimensional domain, representativeness can easily be enhanced by considering difficulty. In a multidimensional domain, however, taking account of difficulty may not be so straightforward. Some people will be better at one kind of performance, some at another. Difficulty would not be a simple ranking.

If the dimensions represent different duty areas, it might be better to stratify each dimension by difficulty, and then to stratify the dimensions. If, on the other hand, there is moderate correlation among the dimensions, it might be acceptable to treat difficulty as comparable across dimensions rather than as meaningful only within each dimension.

The Job Performance Measurement Working Group participants in the discussions pointed out that in the military context specification of job content is to an important degree a matter of policy. Decisions about both the boundaries of jobs and how specific objectives are to be accomplished tend to be prescribed, presumably to bring a measure of uniformity to a large, sprawling institution that is continually replenishing its work force. The point is important to the extent that policy departs from actual job requirements. In any event, the role of policy in defining the domain of job requirements sets an upper limit for the interpretation of test scores. (This is, of course, not unique to the military. Any large employer will have institutionalized job descriptions and performance expectations—once in existence, job analyses tend to become statements of policy. But the system in the military is very highly articulated and probably leaves less room for maneuvering than private-sector researchers are accustomed to.)

This entire discussion has reaffirmed the critical importance of thoughtful job analysis and test content selection. Competency interpretations depend on a high degree of content validity. The committee participants again recommended a statistical sampling model as the most scientifically supportable means of ensuring the representativeness of the test, although the test can only be as good as the job specification on which it is based.

## Test Scoring Strategies

In creating scales, either to show individual differences or to assess level of competency, there are several ways of combining the binary scores on steps to get task scores and several ways to combine task scores to get a total test score. Furthermore, there may be some advantage in creating a profile of test scores for different duty areas as an intermediate level of analysis, as the Army has done, for example, with its common and occupation-specific tasks. Considerations are somewhat different for scoring a task and for combining those task scores to get a test score.

## Scoring a Task

Hands-on tests by necessity include a relatively small number of tasks, but each task has many steps, which are typically scored go/no go. Once pass or fail designations have been assigned for each of the steps in a task, the question becomes how the steps can be combined to get a score on the task, which can then be combined with other task scores to get an overall test score.

For example, suppose that changing a tire is a task on a truck driver's hands-on test. An examinee who cannot operate the jack cannot change the tire. Does this count as a task failure or simply a step failure, with the examiner jacking up the vehicle and the examinee proceeding from there? What penalty is earned by jacking up the vehicle before loosening the lug nuts?

Several scoring models might be considered for combining steps to score a task. The scoring models are here called compensatory, conjunctive, disjunctive, and hybrid. A compensatory model allows an individual to make up for a poor performance on some steps by a good performance on others. A conjunctive model implies that an individual must successfully complete each of the composite steps in turn, a disjunctive model requires success in only one of the components, and a hybrid model is some combination of these elements.

A compensatory scoring scheme for a task involves adding the scores for each step of the task. The step scores can be a simple dichotomy (0,1; go/no go), or they can be weighted. Differential weights allow some steps to count more than others. With sufficiently disparate weights, some steps can completely dominate others.

A purely conjunctive model requires success on every step. A simple example is to require successful completion of each step, to note where in the sequence the first step is failed, and to count the number of preceding steps.

A purely disjunctive model allows success if any one of the steps is achieved. Almost certainly this would apply only to a few of the many steps. For example, one could decide that if the tire is changed, it doesn't matter how well it was done.

A variety of hybrid schemes can now be envisaged. A modified compensatory-conjunctive model would permit the usual compensating scores, provided that one or two critical steps were done correctly. A group of steps could be scored in a compensatory

manner, and then a cut point could be established to turn that group into a 0,1 score depending on whether the performance was above or below the cut The group scores could then be scored in a compensatory fashion.

If the steps in a task form a perfect Guttman scale, then the conjunctive model is identical with the compensatory model. In a perfect Guttman scale, the items (steps) are ordered, with each step harder than those before it, so that success on a given item (step) implies success on all previous steps. But pure Guttman scales are rare. Examinees frequently complete some steps successfully after failing a given step, provided they are allowed to proceed. How to derive a task score must then depend on expert judgment. Automatically adding up the number of successful steps may not be the wisest course, especially if some of the steps are critical.

### Combining Task Scores to Obtain Test Scores

Compensatory, conjunctive, and disjunctive models, which were offered as strategies for scoring steps in a task, are also available for combining tasks to obtain a test score. A compensatory model is usually most appropriate, but the others may sometimes be useful. As an example of a conjunctive strategy, consider the hands-on test for cannon crewman in the Army, which includes several different task groupings, including using the radio, navigating, and using the cannon. Suppose an individual did well on the first two yet poorly on the third. How is that person to be described psychometrically? If it is important for a crewman to know all phases of the job, then rather than summing the scores on all tasks, the groupings could be scored separately, and the poorest score could be taken as the proficiency. By contrast, a disjunctive strategy might involve scoring groups of tasks separately by adding the task scores; the best group score could then be used as the final score.

With a compensatory model, the question of differential weighting arises. Although it would be possible to weight the tasks equally, there might be reason for using weighted scores to reflect a more complex view of the job. The weights might be established by job experts on the basis of a job analysis. This would provide a means for making scores more representative of actual job performance. For example, if the job specification indicates

that simple tasks occur with great frequency, the simple test tasks could be weighted accordingly. If, however, job experts report that the more characteristic feature of a particular job is the necessity for all incumbents to be able to perform a small set of extremely critical tasks, with the remaining tasks being the equivalent of sweeping up, then the tasks representing that critical subset could be very heavily weighted.

Both weighting schemes have policy implications for how competency is evaluated. The decisions may appear to be technical, but in fact they formulate policy. Competency is referenced to the domain of job requirements, but the basis for evaluating competency is the set of observations in the performance measure. Different evaluations of levels of competency would be made depending on the weighting scheme.

Note that the task scores should be made comparable before applying rational weights. If one task has 5 steps and another has 10, then, if the steps are scored dichotomously 0,1 and added, the range of possible scores is twice as great on the second task. A reasonable and simple procedure for putting the tasks on an equal footing would be to divide the task score by the number of its component steps, to get a range from 0 to 1, and then to multiply by some convenient constant like 10 or 100 to get a more comfortable but still equal range of possible scores. Some psychometricians would prefer to standardize the task scores, so that the distribution of task scores had a variance of 1.0 or some other convenient constant. The committee does not advocate equating the empirical variances because that tends to emphasize individual differences rather than emphasizing how much of the task can be done.

From one point of view, it is possible that the outcome of the weighting scheme in terms of evaluating standards may be more illusory than real. If job performance is characterized as a single number, and the observations are summed to obtain a total score, then the correlation between unit-weighted and multiple-weighted scores will be high. Indeed, since negative weights for observations are not reasonable, the correlation may be so high that virtually the same rank order of examinees would obtain under either scaling method.

However, if the performance scores are to be interpreted as measures of competency, with a given test score indicating a certain level of job performance, then the weighting scheme is im-

portant. It should be emphasized that an externally referenced meaning depends on attending to means and standard deviations as well as correlations.

What was said above about correlations of differently weighted scores is still true for externally referenced scores, but the attention shifts away from rankings to mean scores.

The effects of alternative weighting schemes should be investigated in the context of evaluation standards. Is the linkage of job performance and standards affected by the weighting scheme? Obtaining the weights is a laborious process, and to be worthwhile they should have a formative impact on the linking outcomes.

A word of caution is necessary when discussing weighting of tests constructed by stratified random sampling. Differential weights are mainly relevant to tests constructed by purposive sampling of tasks. If a test has been constructed by stratified random sampling of tasks, and if the strata and/or the tasks within strata have been given differential sampling weights as a means of defining the primary performance measure, then the weighting has been done in the sampling and should not be repeated after the test has been formed. Any more elaborate weighting system would tend to mask the central thrust of task sampling in defining the primary score. Different weights would be entirely appropriate for defining alternative measures, as long as they are clearly stated. The notion of representativeness suggests that the task scores be on comparable scales, e.g., all dichotomous (0,1) or all continuous (0-10), and that the task scores be added to get a total score. Subscores for each stratum or group of strata could be entertained, but otherwise equal weighting of the task scores is appropriate. Still, the sampling weights might not be sufficient, in themselves, to reflect extreme differences in task performance. A pilot who cannot land the plane is in deep trouble, regardless of his skill in maneuvering the plane in flight. There might be reason to weight critical tasks differentially even after random sampling of tasks for inclusion on the test.

## Interpretive Scale Anchors

Previous sections have focused on defining the job domain and on selecting and scoring the tasks that comprise a compe-

tency scale. The focus now shifts to interpreting the scale values. One possible approach to providing meaning to the proficiency scores would be to attach descriptive anchors at several regions of the score scale. This would depend on subject-matter experts' being able to agree that a certain region of scores represents the performance of a novice; higher scores would be designated that represent apprentice performance, journeyman, master, and expert. Associated with each label would be a range of behaviors that would be expected of someone with a score in that part of the scale. The National Assessment of Educational Progress (NAEP) uses a similar strategy in explaining levels of reading mastery. Reading is admittedly more nearly unidimensional than performance on most jobs, but the goal has appeal.

Another possible approach to attaching meaning to scores would be to use the five pay grades for first-term enlisted personnel to describe the distribution. A more attractive possibility, if it were feasible, would be to use the already-established skill levels associated with military jobs. (This possibility needs further exploration.) Again one could elicit subject-matter experts' judgments about what kinds of tasks people at each level could be expected to perform.

A third suggestion was to use the Air Force occupational learning difficulty (or their equivalents in the other Services) as the proficiency anchors, recasting them to describe what a job incumbent at each level can do.

The competency discussion group is divided on the question of anchors. Some service representatives feel that anchors amount to multiple cut-points on the performance scale, and they want to avoid anything that suggests performance hurdles or categories of performance. Although the borders between anchors should be viewed as very indistinct, categories have a tendency to be overinterpreted. A person near the top of the apprentice category should be viewed as nearly indistinguishable from a journeyman. The same is true of the border between AFQT Category II and AFQT Category III, but over the years the AFQT mental categories have attained a reality that they do not deserve.

No matter how a performance test is constructed, the process of attaching meaning to the performance scores will involve some evaluation of test performance by subject matter experts. Some

thoughts about how to elicit such judgments are provided in the appendix to this report.

## Taking Account of Experience

One problem that awaits a clearer resolution by the group is the role of experience. One of the factors that might give rise to differential task performance is experience with the particular tasks tested. The incumbent may perform well those tasks done frequently on the job, but not so well those that are not performed daily. If it were the case that all workers in a job could perform competently whatever tasks were a routine part of the job, i.e., if experience were the only variable, then the interpretation of differential test performance would be fairly straightforward.

However, it is more likely that differential performance is the product of a combination of ability and experience differences. For example, the committee was given to believe that the assignment of individuals to tasks within a military occupational specialty tends, at least at some bases, to depend on how well they perform. Top performers, after demonstrating their skill on easier tasks, are placed in more demanding positions. Hence, they are more likely than journeymen performers with equivalent time in service to have practiced many tasks that occur in the sample on a job performance test. For this reason it seems appropriate to define proficiency over all tasks in the sample and use this to infer overall job performance, rather than considering or giving much greater weight to tasks on the test that the individual recently performed.

An alternative view leading to the same conclusion is that the military wants to know if individuals can do the job they are assigned to even if they are not currently practicing all tasks, so that the job measure should include performance across tasks, regardless of recency of practice.

## Scale Comparability

The above discussion of test scoring is concerned with obtaining a competency scale for a single job. Policy makers have to deal with the totality of jobs, so the question of relating competency scales to one another becomes important. Earlier, in discussing the advantages of domain-referenced tests, we noted that for setting minimum standards for each separate job, it would be useful,

after getting meaningful absolute scales of competence for each of several jobs, if the same fixed value (say 40-70) represented journeyman-level performance for all jobs. However, for allocation decisions, as well as for justifying manpower quality requirements, one might want the values assigned to journeyman-level performance in a given job to represent the utility of journeyman-level performance on this job to the overall mission of the Service. The entire question of relating score scales to allow comparisons across jobs requires careful consideration.

## CONCLUSIONS

Although this report adds descriptive detail to the discussions of committee and Job Performance Measurement Working Group members that took place on October 24-25, 1986, and March 13-14, 1987, and perhaps extends the logic of the discussion on some points, it conveys the sense of the meetings. Our joint exploration of the complexities of the subject are encapsulated in the following statements:

1. We have tentatively answered, in the affirmative, the question of whether providing a competency measure for hands-on performance is useful. This competency interpretation might be referenced to the typical tasks that can and cannot be performed at a particular score level.

2. The competency approach seems promising: (a) to link enlistment standards to job performance by providing "meaning" to the score distribution; (b) to improve manpower allocation by increasing the weight of performance factors in balancing the needs of competing occupational specialties; and (c) to provide more credible justification for the Services' quality needs.

3. We have reaffirmed the critical importance of test content selection/content validity. (The committee continues to recommend a statistical sampling model.)

4. We recognize that scoring strategies depend on policy and that the definition of competency will be a product of policy decisions as well as scientific assessment.

5. Therefore, the issues of scale anchors, weighting schemes, scoring strategies, and scale comparability that have been laid out in this report require some hard thought.

# Appendix
# Inferring a Scoring Procedure from Expert Judges

An interesting empirical approach to devising a scoring method for a performance test involves induction from expert judgments. One specific system for eliciting judgments and inferring a scoring system is offered as illustrative of the sort of approach we have in mind, although we want to stress that other procedures may prove more useful in practice.

Suppose that a set of experts were asked to act as judges and assign points to a set of hypothetical individuals who are characterized by their hands-on performance test data in the form of task scores, including completion times when available. One method for collecting such judgments would be to provide 20 to 40 task score profiles (which could be a random sample of real performance profiles based on real job performance measurement), plus one reference profile that would have all items performed correctly (and, when relevant, have them performed with very good times). The reference profile would be treated as representing 100 competency points, and each judge would be asked to assign points (from 0 to 100) to each of the other profiles.* The zero point

*We would want to check that ratings were fairly highly correlated—i.e., that the several judges generated similar rank orderings of the profiles. We would also hope to find that the absolute value scores assigned were similar—thus, for example, we would want not only the same kind of profile to be rated lowest by all judges, but also that the actual scores assigned to that sort of profile be similar (rather than one judge using 50 as the lowest rating while another used 80).

could be defined as absence of performance, i.e., being present but with no activity.

If these judgments can be made reliably, we could then move toward developing competency scoring procedures. By regressing the judgments on the task scores across profiles, we could establish the relative weights that the judges appeared to give to the components. Additional elaborations on this standard "policy-capturing" technique might be considered, because the ways in which items might be combined should not be constrained to a simple additive weighting. It may be useful to ask the judges to verbalize the process they were using in evaluating components. Although their judgments often don't follow their stated rules and usually conform to the multiple regression model, a careful analysis might suggest complexities in the algorithms. Any of a variety of scoring algorithms are open, and we can imagine that one sort of scoring (e.g., compensatory) might be best for one military occupational specialty, whereas another kind of scoring would be better for a different one.

There are a number of advantages to this sort of flexibility. In addition to leaving open the question of what sort of scoring algorithm is possible and the option of varying that algorithm by military occupational specialty, it also provides a sort of final test of whether a set of job performance items has much to do with what experts consider important in a job incumbent.

Of particular relevance is that it uses a metric that could, we hope, have applicability across military occupational specialties—for example, if in one the range of scores associated with a sample of examinees is 60-95, while in another it is 30-90, and in another it is 85-99, that seems potentially useful comparative information. Finally, by inviting the judgments of those who actually supervise people in a given military occupational specialty, this approach leaves open the possibility of responding to situations such as one in which an incumbent might be viewed as quite proficient even if deficient in some areas, because those are areas that, in the experience of the judges, are amply covered by many others in a given group.

---

Various other approaches are possible. See, for example, the several approaches discussed by Sadacca et al. (1986), but note that in that paper "performance constructs" were much broader than the kind of job performance test items considered here.

# References

Brennan, R.L.
1981 *Some Statistical Procedures for Domain-Referenced Testing: A Handbook for Practitioners.* ACT Technical Bulletin No. 38. Iowa City, Iowa: American College Testing Program.

Lennon, R.T.
1956 Assumptions underlying the use of content validity. *Educational and Psychological Measurement* 16:294-304.

Maier, Milton H., and Ann R. Truss
1983 *Original Scaling of ASVAB Forms 5/6/7: What Went Wrong.* CRC 457. Alexandria, Va.: Center for Naval Analyses.

Sadacca, Robert, Maria Veronica de Vera, and Ani S. Di Fazio
1986 Weighting Performance Constructs in Composite Measures of Job Performance. Paper presented at annual meeting of the American Psychological Association, Washington, D.C., August 22-25.

Shavelson, Richard J., and Noreen M. Webb
1981 Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology* 34:133-166.

U.S. Department of Defense
1985 *Defense Manpower Quality.* Report to the House and Senate Committees on Armed Services. Washington, D.C.: U.S. Department of Defense, Office of the Assistant Secretary (Manpower, Installations, and Logistics).

Wigdor, Alexandra K., and Bert F. Green, Jr., eds.
1986 *Assessing the Performance of Enlisted Personnel: Evaluation of a Joint-Service Research Project.* Committee on the Performance of Military Personnel. Washington, D.C.: National Academy Press.