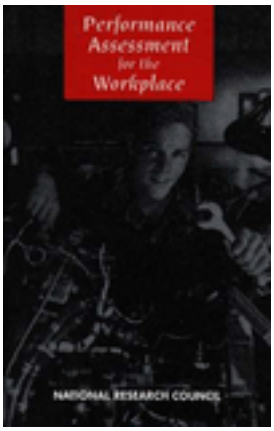


Performance Assessment for the Workplace, Volume II: Technical Issues

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

ISBN: 0-309-59519-3, 344 pages, 6 x 9, (1991)

Performance Assessment for the Workplace

Volume II

Technical Issues

Alexandra K. Wigdor and Bert F. Green, Jr.,
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. 1991

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.

This project was supported by the Office of the Assistant Secretary of Defense (Force Management and Personnel).

Library of Congress Catalog Card No. 91-75424
International Standard Book Number 0-309-04539-8
Additional copies of this report are available from:
National Academy Press
2101 Constitution Avenue N.W.
Washington, D.C. 20418

S376
Printed in the United States of America
First Printing, October 1991
Second Printing, June 1992
Third Printing, September 1992

COMMITTEE ON THE PERFORMANCE OF MILITARY PERSONNEL

* Member, 1983-1985

CONTRIBUTORS

LINDA J. ALLRED, Department of Psychology, East Carolina University
STEPHEN B. DUNBAR, Lindquist Center, University of Iowa
ROBERT GLASER, Learning Research and Development Center, University of Pittsburgh
SHERRIE GOTT, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas
LINDA S. GOTTFREDSON, College of Education, University of Delaware
BERT F. GREEN, JR., Department of Psychology, Johns Hopkins University
RICHARD M. JAEGER, School of Education, University of North Carolina, Greensboro
SALLIE KELLER-McNULTY, Department of Statistics, Kansas State University
ALAN LESGOLD, Learning Research and Development Center, University of Pittsburgh
ROBERT L. LINN, School of Education, University of Colorado, Boulder
PAUL R. SACKETT, Industrial Relations Center, University of Minnesota
RICHARD J. SHAVELSON, Graduate School of Education, University of California, Santa Barbara
FREDERICK D. SMITH, Department of Psychology, Pennsylvania State University
ALEXANDRA K. WIGDOR, National Research Council, Washington, D.C.

Preface

In 1981, the four military Services, in an effort to improve their control over manpower quality in the enlisted ranks, launched a pioneering research program to develop measures of job performance so that, for the first time, enlistment standards could be linked to performance on the job. The Joint-Service Job Performance Measurement/Enlistment Standards (JPM) Project, as it is called, is being carried out by each Service under the overall direction and coordination of the Office of the Assistant Secretary of Defense for Force Management and Personnel. In 1983, the Committee on the Performance of Military Personnel was established within the National Research Council to act as an independent adviser to the Department of Defense on the JPM Project. At the request of its sponsors, the committee has given attention to the potential usefulness of the JPM research for personnel decisions and manpower management.

The Department of Defense decided to undertake the JPM Project as a result of difficulties caused by a technical error in scoring the test that is used throughout the military to determine enlistment eligibility: the Armed Services Vocational Aptitude Test Battery (ASVAB). The error had the effect of inflating test scores, so that approximately 250,000 young men and women were inducted between 1976 and 1980 who did not actually meet the entrance standards. The issue was complicated by concerns about the success of the all-volunteer force, because some of the Services had been having trouble in the late 1970s meeting recruiting goals. As a consequence

of the test misnorming, policy makers in both Congress and the Department of Defense became interested in establishing the relationship of the ASVAB to actual job performance. The JPM Project was the Department's response to those concerns.

The first phase of the JPM Project was to determine if accurate, valid, and reliable measures could be developed that are representative of job performance and to determine how well the current enlistment procedures, including the ASVAB, predict these approximations of job performance. The second phase, now well under way, is to develop ways to set enlistment standards using the new job performance data. More specifically, the Department is exploring the use of cost/performance trade-off models to provide the standards-setting process with a more solid empirical foundation.

The primary focus of the committee in the early years of the JPM Project was the overall research design and the development of instruments to measure job performance. Later, the focus turned to problems in hands-on test administration, controlling for unreliability of measurement, the relationships among the various new performance measures, and extending the research findings to a larger set of military jobs. In order to place the research in context, the committee also learned about military entrance processing, entry-level jobs in the military, technical training, and the general outlines of how entrance standards are set. Committee members made a series of site visits to Army, Navy, and Air Force bases to see enlisted personnel at work, to talk to their supervisors about the content of entry-level jobs, and to observe test administration procedures. Subgroups of committee members made a number of trips to military personnel research laboratories to gather information. To facilitate an interchange of ideas, the committee invited JPM Project scientists as well as other experts to explore solutions to specific technical problems in a series of workshops. And, as supplements to its activities, the committee has called on outside experts to prepare background materials on various aspects of the issues involved.

Since 1983, a series of reports has been delivered periodically to the Department of Defense on various aspects of the JPM Project. The final report, which is companion to this volume, summarizes what the committee learned from analyzing the JPM experience. It begins with a historical overview of the criterion problem and a discussion of the conceptual approach and general research design of the project. It then looks closely at specific issues: the development of performance measures; sampling, logistical, and standardization problems; evaluating the quality of performance measurements in terms of reliability and content representativeness; the relationship between test scores and criterion measures; and management of human resources. The committee hopes that the insights and information contained therein will be of value to an audience wider than the military services, including policy makers, members of the testing community, employers

concerned with performance assessment, and, given the new currency of performance assessment in the education arena, to the many school officials, educators, and policy makers involved in education reform.

This volume contains some of the most valuable papers that were prepared for the committee. With them the authors helped the committee and the JPM research scientists think through the technical challenges raised by attempts to develop criterion measures for a sample of jobs that could be made meaningful to the universe of jobs in the services. Some focus on approaches to performance measurement and analysis of the job performance data; others deal with broader issues involved in comparing multiple measures and generalizing from a small sample of jobs. Taken together, they provide those interested in the technical details of the JPM Project a closer look at some of the problems, challenges, and possible solutions. We sound their themes in the paragraphs that follow.

Robert Glaser, Alan Lesgold, and Sherrie Gott, in a paper that looks to the next generation of performance measurement, discuss the methodology needed to measure the cognitive aspects of job performance. The large number of highly technical jobs and the short periods of enlistment in which both training and useful performance must take place make the problem especially severe for the services. Cognitive psychology has produced a variety of methods that can be sources of a new set of measurement methodologies; the authors' application of these techniques in developing a cognitive task analysis procedure for technical occupations in the Air Force is the basis for their conclusions. They present a cognitive account of the components of skill, discuss the specific measurement procedures employed, and consider which aspects of measurement in the services can best use these approaches.

The measurement method of greatest interest in the JPM Project is the work sample. Frederick D. Smith presents a review of the work sample literature, with particular attention paid to the theory underlying work sample testing, the use of work samples as criterion measures, the adverse impact of work samples, and measurement issues associated with such tests. Considering both their advantages and disadvantages, he concludes that the research concerning work sample testing suggests that they can produce high predictive validities, and that when used as criteria they compare favorably with supervisor ratings and productivity measures.

In a paper drafted on behalf of the committee, we discuss the meaning of assessing competency or job mastery and consider ways of establishing such interpretations and using the results. We suggest that, to effectively communicate information about the performance of enlisted personnel and the implications of changing standards, the scoring scale of the job performance tests needs to be given some sort of absolute meaning. Policy makers would be better able to make informed judgments about what distribution

of "quality" in the recruit cohort 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. We illustrate this argument by analyzing a simple model for setting entrance standards.

Whereas many of the papers in the volume are concerned with developing more adequate measures of job performance, Linda S. Gottfredson explores strategies for evaluating alternative kinds of criterion measures. Today's challenge, she argues, is to develop procedures for comparing the relative utility of alternative measures for a given purpose. Gottfredson presents interesting suggestions for assessing the types and degrees of similarity and differences among criterion measures. Although she focuses on evaluating job performance measures in their role as criteria in developing personnel selection procedures, it has more general applicability.

Stephen B. Dunbar and Robert L. Linn provide an overview of standard procedures used to adjust correlations and regression parameters for the effects of selection, commonly referred to as corrections for range restriction. Technical issues related to the accuracy of these adjustments are considered, especially where they are likely to have implications for the types of adjustment procedures appropriate for large-scale predictive validity studies of an aptitude battery like the ASVAB. The authors conclude with a discussion of issues related to the implementation of a set of adjustment procedures for validation studies in the military, where the choice of the reference population, choice of selection variables for making adjustments, and choice of an analytical procedure all have important consequences for the assessment of the validity of the ASVAB for predicting performance on the job.

Linda J. Allred considers alternatives to the validity coefficient for reporting the relationship between test scores and performance. The validity coefficient indicates the overall strength of the test-criterion relationship for the groups being studied, but its meaning is obscure to a nontechnical audience. Using several sets of hypothetical data, Allred illustrates various display methods, including the scatter plot, the box-and-whisker plot, expectancy methods (chart, table, and plot), and the frequency table, and describes their strengths and weaknesses.

Richard J. Shavelson lays out a statistical theory of the multifaceted sources of error in a behavioral measurement. Called generalizability (G) theory, the theory has heretofore been applied to traditional measurements such as aptitude and achievement tests. Shavelson provides an example of how G theory can be applied to military job performance measurements, using hypothetical data, as well as specific applications of the theory, chosen to highlight the theory's flexibility in modeling a wide range of measurements.

Richard M. Jaeger and Sallie Keller-McNulty address three problems

associated with the use of hands-on tests of job performance. The first concerns methods for setting standards of minimally acceptable performance on the tests. The second involves procedures for eliciting and combining judgments of the values of enlistees' behaviors on military job performance tests. The third concerns procedures for using enlistees' predicted job performance test scores and judged values associated with those test scores in classifying enlistees into military occupational specialties. The first, for which there is the greatest research available, is discussed is considerable detail; discussion of the second is comparatively brief; and discussion of the third is illustrative rather than definitive.

Paul R. Sackett considers approaches to extending validity findings and empirically based predictor cutoffs beyond the 27 jobs chosen for inclusion in the JPM Project to the universe of military occupational specialties. The purpose for which job analysis is being done or for which jobs are being compared is often ignored, and the choice of the job descriptor has an important impact on decisions about job similarity. No single approach is recommended; rather, a number of possibilities are examined.

To all these authors the committee is grateful for turning their knowledge and experience to a number of novel and exceedingly difficult technical issues confronting all of those who would address the criterion problem seriously. They have enriched the advice that the committee provides to the Department of Defense, applying the sciences of psychometrics, testing and performance measurement, and industrial psychology to the problems raised by the JPM Project. We commend this volume to those who wish to expand their understanding of the issues, challenges, and advances generated by one of the largest and most important studies of job performance on record.

BERT F. GREEN, JR., CHAIR

ALEXANDRA K. WIGDOR, STUDY DIRECTOR

COMMITTEE ON THE PERFORMANCE OF MILITARY PERSONNEL

Contents

Implications of Cognitive Psychology for Measuring Job Performance

Robert Glaser, Alan Lesgold, and Sherrie Gott

INTRODUCTION

In comparison to a well-developed technology for aptitude measurement and selection testing, the measurement of learned occupational proficiency is underdeveloped. The problem is especially severe for the Services because of the many highly technical jobs involved and the short periods of enlistment in which both training and useful performance must take place. To increase the effectiveness of both formal training and on-the-job learning, we need forms of assessment that provide clear indicators of the content and reliability of new knowledge. Since many of the military's jobs have a major cognitive component, the needed measurement methodology must be able to deal with cognitive skills.

Fundamentally, the measurement of job performance should be driven by modern cognitive theory that conceives of learning as the acquisition of structures of integrated conceptual and procedural knowledge. We now realize that someone who has learned the concepts and skills of a subject matter has acquired a large collection of schematic knowledge structures. These structures enable understanding of the relationships necessary for skilled performance. We also know that someone who has learned to solve problems and to be skillful in a job domain has acquired a set of cognitive procedures attached to

This paper has not been cleared by the Air Force Human Resources Laboratory and does not necessarily reflect their views.

knowledge structures, enabling actions that influence goal setting, planning, procedural skill, flexibility, and learning from further experience.

At various stages of learning there exist different integrations of knowledge, different degrees of procedural skill, differences in rapid memory access, and differences in the mental representations of tasks to be performed. Proficiency measurement, then, must be based on the assessment of these knowledge structures, information processing procedures, and mental representations. Advancing expertise or possible impasses in the course of learning will be signaled by cognitive differences of these types.

The usual forms of achievement test scores generally do not provide the level of detail necessary for making appropriate instructional decisions. Sources of difficulty need to be identified that are diagnostic of problems in learning and performance. An array of subject matter subtests differing in difficulty is not enough. Tests should permit trainees to demonstrate the limits of their knowledge and the degrees of their expertise. The construction of tests that are diagnostic of different levels of competence in subject matter fields is a difficult task, but recent developments, including cognitive task analysis and research on the functional differences between experts and novices in a field, provide a good starting point for a theory to underpin proficiency measurement.

Until recently the field of psychological measurement has proceeded with primary emphasis on the statistical part of the measurement task, assuming that both predictor and criterion variables can be generated through rational and behavioral analysis and perhaps some intuitions about cognitive processing. This has worked quite well when the fundamental criteria are truly behavioral, where the valued capability is a specific behavior in response to a specific type of event. However, when the fundamental performance of value is cognitive, as in diagnosing an engine failure or selecting a battlefield tactic to match a determined strategy, more is needed. The problem is particularly apparent if the true goal is *readiness* for a situation that cannot be simulated entirely or if it is to decide what specific additional training is required to assure readiness.

We are impressed by the fact that much of the technology of testing has been designed to occur after test items are constructed. The analysis of item difficulty, discrimination indices, scaling and norming procedures, and the analysis of test dimensions or factors take place after the item is written. In contrast, we suggest that more theory is required before and during item design. We must use what we know about the cognitive properties of acquired proficiency and the mental processes and structures that develop as individuals acquire job skills. The nature of acquired competence and the indicators that might signal difficulties in learning are not apparent from a curriculum analysis of the facts and algorithms being taught.

Proficiency measurement should be designed to assess not only algorithmic knowledge but also cognitive strategies, mental models, and knowledge

organization. It should be cast in terms of levels of acquisition and should produce not only assessments of job capability but also qualitative indicators of needed further training or remediation. In this regard, cognitive psychology has produced a variety of methods that can be sources of a new set of measurement methodologies. During the past 2 years, we have been applying these techniques in developing a cognitive task analysis procedure for technical occupations in the Air Force. The conclusions in this paper are based on our experience in this project.

The three sections that follow present a cognitive account of the components of skill, discuss the specific measurement procedures we have employed, and then consider which aspects of measurement in the Services can best use these approaches.

COGNITIVE COMPONENTS OF SKILL

There are three essential elements in cognitive tasks. These are

- The *contents* of technical skills: the procedures of which they are composed.
- The *context* in which technical skills are exercised: the declarative knowledge needed to assure that skill is applied appropriately and with successful effect.
- The *mental models* or *intermediate representations* that serve as an interface between procedural and declarative knowledge. These three essential aspects of job proficiency are emphasized throughout this paper.

Procedural Content

A starting point for specifying the procedural content of a task is the GOMS model proposed by Card et al. (1983) in their studies of the acquisition of skill. This model has been further elaborated in work on formal procedures for representing the complexity of machine interfaces for users (Kieras and Polson, 1982). The GOMS model splits technical knowledge into four components: goals, operators, methods, and selection rules. We have adopted a similar split with a few differences from GOMS. Each component is not only a subset of the total knowledge a technical expert must have, but is also a reminder to consider certain issues in attempting to understand the expertise.

Goal Structure

Any task can be represented as a hierarchy of subgoals, and experts usually think of tasks this way. Generally, such goal structures become

very elaborated for complex tasks. The overall goal is decomposed into several nearly independent subgoals, and they in turn are subdivided repeatedly. In many cases, the particular subdivision of a goal depends on tests that are performed and the decisions that are made as part of the procedure that accomplishes the goal. Something like the following example[1] is quite common:

```
TO TRAVEL-TO :X
SUBGOAL CHECK-OUT-POSSIBILITIES
IF :X WITHIN 50 MILES THEN SUBGOAL DRIVE-TO :X
ELSE SUBGOAL FLY-TO :X
END
```

Because of this sort of contingent branching, an overall subgoal structure for a particular goal may not exist explicitly. Rather, it may be assembled as the goal is being achieved—it is implicit in part.

Without substantial prior experience, it can be difficult to separate a complex task into subgoals that are readily achieved in a coherent manner and that do not interact. A novice, even if intelligent, may separate a task into pieces that cannot be done independently. Consider the following example of a goal structure:

```
TO HAWAII-VACATION
SUBGOAL GET-TICKETS
SUBGOAL RESERVE-CAR
SUBGOAL RESERVE-HOTEL
SUBGOAL GET-THERE
SUBGOAL PAINT-TOWN-RED
END
```

Suppose that a person adopts this goal structure. Some problems could develop. For example, he may have a budget limit. If in solving the GET-TICKETS subgoal he uses up too much money, he then will not be able to solve other subgoals. We say the subgoals interact. Also, it is possible that a package deal can solve all three initial subgoals, so subdividing them as shown is unnatural and may divert the novice from a successful approach.

There is another form of novice goal setting that is almost the opposite. This is the use of subgoals that are defined by the methods the novice knows rather than the overall goal to be achieved. Again, this is often a case of intelligent behavior by those not completely trained. For example, a rather inexperienced engineer given the task of designing a conditioned power source for a large computer was heard to say: "What you need is a

[1] The example is stated using some of the formalisms of LOGO.

UPS [uninterruptible power source] without the batteries." The engineer happened to have learned about uninterruptible power sources and knew that such systems provided clean power. However, he knew nothing about how, in general, clean power can economically be provided. By looking only at designs for UPS systems, he missed some cost-effective designs that work fine *except* when battery back-up is also required.

In analyzing a technical specialty, it is necessary to establish what goal structures are held by experts. Sessions in which the expert describes how a task is carried out are very helpful for this purpose, and we have made heavy use of them. Expert-novice comparisons are also helpful. Also, information about novice goal structures can sometimes reveal training problems that could be corrected by specifically targeted instruction.

Basic and Prerequisite Abilities (Operators)

Card et al. (1983) spoke of the basic operators for a task domain, borrowing from the earlier Newell and Simon (1972) approach of specifying elementary information processes from which more complex procedures would be composed. We have unpacked this idea a bit. For any given technical specialty, there are certain basic capabilities that the novice is assumed to have prior to beginning training. For example, one might assume that the ability to use a ruler (at least with partial success) might be a prerequisite for work as an engine mechanic. Ruler use is hardly a primitive mental operation in any general sense, but with respect to subsequent training in an engine specialty, it could be considered as such. Similarly, the ability to torque a bolt correctly might be thought of as a basic entering ability for new three-level airmen coming to their first operational assignments. What is important is that every training approach makes these assumptions concerning prerequisites and that some such assumptions are incorrect. Often, for example, a training approach will assume a highly automated capability but will only pretest for the bare presence of that capability.

In essence, we are asserting that the components of a skill should be subdivided into two relevant parts: those that are prerequisite to acquisition of a particular level of skill and those that are part of the target level. What is prerequisite at one level may be a target component at a lower level. In doing a cognitive task analysis, one must examine the performance of successful novices to determine what the real skill prerequisites are and what new procedures are being acquired. Such an analysis must take account not only of nominal capability but also of the speed and efficiency of prerequisite performance capabilities. Care must be taken to avoid declaring too many skills as prerequisites. A common fault of educational and training systems is to declare many aspects of a skill to be prerequisite and then

bemoan the lack of adequate instruction of those prerequisites in lower level schools.

Procedures (Methods)

At the core of any task analysis is an analysis of the procedures that are carried out in doing the task. This remains the case with cognitive task analyses. A major part of our current work on cognitive task analysis of avionics equipment repair skills is to catalog the procedures that an airman must know in order to perform tasks within the job specialty being studied. This is done in a variety of ways and is probably the aspect of cognitive task analysis that is closest to traditional rational task analysis approaches. We examine technical orders, expert and novice descriptions of tasks, and other similar data. While the resulting procedural descriptions are likely to be similar to those achieved by earlier approaches, they are distinguished by the following new components: (1) we are attending explicitly to the enabling conditions, such as conceptual support (see below), for successful procedure execution, and (2) separate attention is paid to goal structures and selection rules. Further, we use a variety of techniques to verify our analyses empirically.

Procedural descriptions of cognitive tasks will tend to emphasize the domain-specific aspects of performance. However, it will sometimes be appropriate to include certain self-regulatory skills of a more general character in the analyses. It is critical to avoid basing cognitive analyses on the ability of people with strong self-regulatory and other meta-cognitive skills to handle many novel tasks. When a task requires these more general skills in addition to easily trainable specific capabilities, then this should be noted. In general, it can be assumed that a continual supply of trainees with strong self-regulatory skills or other high levels of aptitude cannot be guaranteed, and such skills are not quickly taught.

Selection Rules

The progress of modern psychology has been marked by a slow movement from concern with stimulus-response mappings to a concern with mappings between mental events (including both perceptions and the products of prior mental activity) and mental operations or physical actions. To the extent that they stuck to the earlier methodologies evolved from stimulus-response approaches, trainers knew only that certain physical responses must be tied to certain stimuli. They had only indirect ability to teach by rewarding correct responses and punishing errors. Now, new methodologies are being developed for verifying mappings between internal (mental) events and mental operations. While they still have pitfalls for the unwary, they make possible an interpretation of tasks that comes closer to being useful for instruction.

In particular, we know that the knowledge of experts is highly procedural. Facts, routines, and job concepts are bound to rules for their application, and to conditions under which this knowledge is useful. As indicated, the functional knowledge of experts is related strongly to their knowledge of the goal structure of a problem. Experts and novices may be equally competent at recalling small specific items of domain-related information, but proficient people are much better at relating these events in cause-and-effect sequences that reflect the goal structures of task performance and problem solution. When we assume that some training can be accomplished by telling people things, we are, in essence, assuming that what the student really needs to know are procedures and selection rules for deciding when to invoke those procedures.

In conducting a cognitive task analysis, it is important to attend specifically to a trainee's knowledge of the conditions under which specific procedures should be performed. Combined with goal structure knowledge, selection rules are an important part of what is cognitive about cognitive task analyses.

Conceptual Knowledge

In performing cognitive task analyses, it is important to consider how deep or superficial the knowledge and performance are. Many skills have the property that they can be learned either in a relatively rote manner or can be heavily supported by conceptual knowledge. For example, one can perform addition without understanding the nature of the number system, so long as one knows the exact algorithm needed. Similarly, one can repair electronics hardware without deep electronics knowledge, so long as the diagnostic software that tells one which board to swap can completely handle the fault at hand. However, it appears that the ability to handle unpredicted problems, which is a form of transfer, depends on conceptual support for procedural knowledge.

The difficult issue (as with automation of skill) will be the separation of conceptual knowledge evidenced by experts and quick learners into that which is a mere correlate of their experience and that which is necessary to their experience. This is still very much an issue of basic research, but one on which some good starting points can be specified. The sections below detail three types of conceptual supporting knowledge to which cognitive task analyses should be sensitive.

Task Structure

Some, but not all, experts retain detailed knowledge of the structure of tasks that they perform. We usually call these people the good teachers. Part of their knowledge results from having explicit, rather than implicit, goal

structures. A serious question to be addressed in any analysis is whether possession of certain explicit task structure knowledge is necessary to successful skill acquisition. A corollary issue is whether those who do not possess or have trouble acquiring such explicit knowledge tend to acquire their skill in a different manner from those who are more "high verbal." This is part of the "basic research" aspect of cognitive analyses at this time, because general domain-independent answers to such questions have not yet been established.

Background Knowledge

The same set of questions applies to a second form of conceptual support knowledge, namely background knowledge. In electronics troubleshooting, for example, it is conceivable that a bright person could perform many (but not all) of the tasks much of the time by simply following the directions in the printed technical orders. While this might take a long time, it seems at least possible. What is added to performance capability by knowing about basic electrical laws or about how solid state devices of various types work? The cognitive task analyst must attempt to determine the role played by such knowledge in successful performance. The knowledge to be examined includes scientific laws and principles as well as more informal background, such as crude mental models and metaphors for processes that are directly relevant to job tasks (e.g., what happens when a circuit is shorted).

Context of Use

Related to this second type of conceptual knowledge is contextual knowledge. For example, a novice jet engine technician whom we studied resorted to knowledge of air flow through a jet engine to answer some of our questions about engine function. The immediate question this raises for us is whether successful module replacement (a role that includes no diagnosis) depends on any knowledge of how a jet plane works, what the pilot sees or experiences, or what other roles are involved in servicing a plane besides module replacement. Again, while we have found that our better subjects know a lot about all of these contexts, the chore of the cognitive task analyst is to determine the extent to which this knowledge is necessary for successful skill development (or to one success track in skill development).

Critical Mental Models

Expertise is generally guided by several kinds of mental models. One kind is the model of the problem space, as one might see in a chess player, who has a rich representation of the board positions to which he can anchor various interpretations and planned actions. A related kind of mental model

is the model of a critical referent domain, such as the model of the patient's anatomy that is maintained by a radiologist. Such a model is not identical to the problem space but rather is an important projection of the problem. Finally, there are critical device and component models that guide understanding and performance, such as the electronics technician's model of a capacitor or of a filtered power supply. Each type of model can be critical to problem solving, and the task of the cognitive task analyst is to discover which models of any type play an important role in expertise.

There are several ways in which this can be done. For some domains, such as electronics troubleshooting, there is an existing literature because cognitive scientists have used the domain to study expertise, have been designing tutoring systems for the domain, and have been building expert systems to supplement human expertise in the domain. Another approach is to ask experts and less-expert workers to think out loud while solving problems in a domain. In our laboratories, we have done this, for example, with radiologists (Lesgold, 1984; Lesgold et al., 1985). Analysis of the verbal protocols from these physicians (taken while they made film diagnoses) led us to a clearer understanding of the specifics of the mental model of the patient's anatomy that seems to be the focal point of much expert reasoning in this domain. In our more recent work, we have developed much more objective and practical approaches to getting and analyzing protocol data.

A critical finding coming out of the radiology work is that experts have pre-existing schemata that are triggered early in a diagnosis. These schemata tune the patient-anatomy representation and also pose a series of questions that the expert addresses while trying to fit the schema to the specific case. In a sense, then, each disease schema can be thought of as a prototype representation of patient anatomy, and the expert's task is to integrate the schema, the knowledge he has of the patient, and the features of the specific film he is examining. Such schema- and representation-driven processing also occurs in such domains as electronics. For example, good technicians have, among many others, a broad schema for connection failures. When the technical orders fail to provide a basis for a diagnosis, or when computer-based diagnosis fails, schemata such as the connection-failure schema are applied, if possible.

Associated with such schemata in some cases are series of tests that must be performed to either verify or to disprove the fit of schema to fault. In the case of electronics, for example, simply knowing that connection failures cause bizarre and difficult-to-diagnose failures is insufficient. The expert also must be able to build a plan for finding the specific connection that has failed in the piece of gear he is troubleshooting. Detailed study of the technical orders in collaboration with a subject matter expert can be helpful in developing an understanding of the critical schemata for a domain and the content of those schemata. The subject matter expert will often describe the schemata he would apply in a given case if adequately prompted.

Knowledge Engineering is One Aspect of Cognitive Task Analysis

Subject matter experts do not develop clear accounts of expertise on their own. This is why a whole new job category, knowledge engineer, has arisen in the field of expert systems development. Knowledge engineers have, mostly independently of psychologists, developed their own methodology for extracting knowledge from subject matter experts. For example, a standard, perhaps *the* standard, method is to have the expert critique the performance of a novice (sometimes the knowledge engineer plays the role of novice). The idea is that the knowledge engineer can follow the reasoning of the novice but not that of the expert, yet. By asking the expert to critique the level of performance he understands, the knowledge engineer is essentially asking for a repair of his own understanding. Carrying out this process iteratively is a very effective skill for learning the domain and also an appropriate tactic for a cognitive task analyst.

Our own exploratory efforts in cognitive task analysis lead us to make two important cautionary statements about this aspect of cognitive task analysis in particular and the entire approach in general:

- A cognitive task analysis stands or falls partly on the level of expertise in the target domain that is assimilated by the analysis team. This is not a chore for dilettantes.
- Subject matter experts cannot do cognitive task analyses on their own. Because expertise is largely automated, they do not always realize all of the knowledge that goes into their own thinking.

Levels of Acquisition

In addition to specifying the kinds of knowledge needed to do a job well, the cognitive task analyst attempts to understand the level of acquisition that is required. It is important to recognize that knowledge is initially precarious, requiring conscious attention, and relatively verbal. With practice, aspects of skill become sufficiently automated to permit overall performance that is facile and precise. John Anderson (1982) has proposed a theory of skill acquisition that builds upon earlier work on the nature of human thinking and memory (Anderson, 1976). This theory, which elaborates earlier work that was driven by concerns with perceptual-motor skills (Fitts, 1964), provides a useful starting point for analyses of technical skill domains, and it has been incorporated into our approach (Lesgold, 1986; Lesgold and Perfetti, 1978). A number of other researchers (many cited in Anderson, 1982) have anticipated aspects of the approach.

There are qualitative stages in the course of learning a skill. Initially, a skill is heavily guided by declarative (verbal) knowledge. We follow formulae

that we have been told. For example, new drivers will often verbally rehearse the steps involved in starting a car on a hill, while experts seem to do the right thing without consciously thinking about it. A second stage occurs when knowledge has been proceduralized, when it has become automatic. Finally, the knowledge becomes more flexible and at the same time more specific; it is tuned to the range of situations in which it must be applied.

Declarative Knowledge

The measurement of declarative knowledge about technical skills is perhaps the chore that traditional test items handle best. Measures that involve telling how a task is performed are ideal for this purpose. Declarative knowledge can, of course, be partitioned into categories such as goal structure, procedures, selection rules, and conceptually supporting information. When analyzing verbal protocols, it is also important to distinguish between verbal protocol content that provides a trace of declarative knowledge of a task and content that reveals the mental representation(s) that guide performance even after it is automated.

A significant aspect of a skill is the ability to maintain mental representations of the task situation that support performance. For example, in diagnosing a failure of a complex electronic system, a technician sometimes has to have a model of what that system is doing and how information and/or current flows. Such a model, because it is anchored in permanent memory, helps the performer overcome temporary memory limits brought on by a heavy job load and helps preserve memory for the current status of a complex goal structure.

Skill Automation

While it seems essential to successful overall performance, the automation of process components of skill has been difficult to measure adequately. We have used some response speed measures, but speed measures depend heavily on norms for their interpretation, and such norms are seldom available or easily established. Also, and perhaps more important, speed is a characteristic outcome of increasing performance facility, whether or not it is a *cause* of that facility. This issue of causal relationships between subprocessing automaticity and overall performance has been addressed elsewhere (e.g., Lesgold et al., 1985), but the only solution proposed is longitudinal study, which is incompatible with most military measurement needs. (There may be new work soon, from researchers such as Walter Schneider (1985), on this problem.)

One possible approach is to embed the concern over automaticity into all tasks in a test battery, continually watching for evidence of the extent of

skill automaticity and of the extent to which skill shortcomings seem related to lack of automaticity. For example, one can determine the extent to which various external cues, such as diagrams of jet engine layouts, are essential to task performance. Also, one can observe whether the goal structure of a subject exists independent of technical orders that can be referred to. A subject who tells us how to do a task without reference to technical orders must have much of his or her knowledge automated.

Skill Refinement

Finally, one can also look for specific evidence that skills have been refined to the point where there is sensitivity to small but critical situational differences. Flexibility of skilled performance is observed for rapid access to changed representations of a situation given relevant new data.

Summary

Our analysis of job performance highlights the following components of skill:

- knowledge of the goal structure of a task
- skill and knowledge prerequisites for successive levels of performance
- procedural skills and the rules for deciding when to apply them
- conceptual knowledge and metaphors that support performance
- mental models and task representations
- levels of learning, from declarative to proceduralized knowledge, from rigid algorithms to flexible strategies

In general, this approach to job performance is intended to avoid using performance correlates as the basic units of analysis, to instead base measurement and evaluation on an analysis of the specific cognitive procedures and conceptual representations that produce successful performance.

METHODS FOR COGNITIVE TASK ANALYSIS MEASUREMENT

Appropriate methods for cognitive task analysis and for extending cognitive task analysis to the creation of diagnostic test items are continually evolving as more is learned about expertise and the acquisition of proficiency. In this section, we survey several current methods in order to provide a sense of what is possible. The list is by no means exhaustive. It should also be noted that we do not assert that existing military personnel data should be ignored. On the contrary, existing occupational survey data in the Air Force, for example, were extremely useful in focusing our attention on problem areas that merited the expensive cognitive procedures we were developing.

While it is important to be aware of the data already being collected, it is also important to understand their limitations with respect to cognitive analyses. Much of the data are gathered in the course of selecting recruits for specific billets. An instrument might be very effective at picking the right people to be taught a job without being particularly good at specifying how those people who make the cutoff for selection will differ in either their ability to learn or their post-training performance. In essence, when looking at the incumbents within a specific military job specialty, one is looking at a group whose members are chosen because they are classified in the same manner by the available selection tests—as appropriate trainees. There may be further information in the test data, but the tests were designed to serve only as selection instruments. A major purpose for cognitive analyses is to go further than this. The purpose is to identify the kinds of skill and knowledge that must be acquired in school and on-the-job experience, that are basic to the development of job competence. Assessment of these basic skills at the end of training or during the first term on the job might also further inform the selection process.

Procedure Ordering Tasks

For procedural tasks, perhaps the most obvious question one can ask about performance is whether it is carried out correctly. However, it is not always easy to actually have the target performances carried out in a testing situation, nor is it clear how such performances should be scored. Stopping short of actual performance of the target task, one can either ask the subject to tell how the task is done or to reproduce the steps in the task and their ordering from memory, or one can develop tasks in which the steps of the task are displayed to the subject, who must put them into the correct order. This latter approach has the advantage that it is less dependent on the verbal communication and memory skills. However, it still leaves the scoring problem.

When the experimenter tells the subject which steps are included and the subject needs only to order those steps, it appears as if all the hard work is being done by the experimenter. After all, isn't the problem remembering what to do in the first place? As it turns out, there are many cases, perhaps the cases of greatest interest since they represent the harder, less uniformly mastered skill components, in which specifying the order of steps is quite difficult even if the possible steps are shown.

From another point of view, sequencing of procedures should not be a problem, because technical orders are available that specify exactly how a complex procedure should be done, and military personnel are supposed to follow those orders exactly. Unfortunately, this is not always possible and certainly not always optimal. In assembly/disassembly tasks, the technical orders generally assume a completely disassembled device to start with, but in

practice devices are often only partially disassembled, and the order of steps to reassemble them as shown in the technical orders may not work. For example, if one crosses out the steps already carried out and simply does the rest in the order listed, problems can arise. Sometimes, an earlier step in a technical order cannot physically be carried out if a later step has already been done (i.e., you may have to remove one part to reattach another). On other occasions, a particular ordering is physically possible but will not preserve calibrations that are necessary to overall device function. Thus, there are cases in which the ability to adapt the order of steps in a procedure to specific circumstances not anticipated in training or in work aids is a good indicator of depth of procedural knowledge and procedure adaptability.

These cases seem to involve (1) the possibility that the steps in the procedure *could* be carried out in several different orders and (2) constraints on ordering that would not be regulated by feedback the subject receives in the course of actually carrying out the procedure (that is, incorrect orders might not result in immediately observable consequences). In our analyses of jet engine mechanics, for example, we found that specifying the order of steps in carrying out certain rigging (calibration) operations was not something every subject did well. Further, there were systematic relationships between error rates and the ratings subjects received from their supervisors for job effectiveness. In addition, the errors subjects made could, in fact, be neatly classified as conceptual errors or procedural errors. Procedural errors were errors that would have led to an impasse in the course of doing the assembly. Conceptual errors would not have blocked the assembly, but the plane would not have functioned properly afterwards.

Sorting Tasks

Sorting tasks are an important exploratory tool for cognitive task analysis. Indeed, under such names as "Q sort," they have a long history of accepted use in a number of areas of psychology. Initially, the method was used in areas such as social psychology, personality psychology, and advertising research. However, in recent years, the approach has also been used widely in studies of the organization of memory and of expertise (cf. Chi et al., 1982).

The basic theory underlying the approach is that concepts are defined in the mind by a set of characteristic features.[2] When asked to sort pictures or

[2] This account is greatly simplified in order to convey the essence of the approach. In fact. psychologists studying concept formation argue over whether all concepts are characterized this way, whether the features for a concept include defining features shared by all instances as well as typical or characteristic features that are not universal over all instances, etc. These arguments do not affect the validity of the sorting procedure as an exploratory method, and it has been repeatedly demonstrated to be effective in elucidating differences between skill levels in domains of expertise.

words into piles of things that "go together," these features are the most available information in the subject's memory that can be used for such a purpose, so they are used. However, there are many features associated with most concepts, and the subject in a sorting task must make some decisions about which are the most appropriate basis for partitioning the items into separate groups. It is these decisions that seem to vary with expertise.

The general method involves having subjects place in separate piles on a table top the various things being sorted, usually cards with words, phrases, or pictures on them. A record is made of which items ended up in which piles. This is easily done if there are code numbers on the backs of the cards. For large-scale administration, bar coding the cards and scoring through use of a portable data entry device with a bar code reader would be very straightforward. Sometimes, after doing an initial sort, subjects are asked to decompose their piles into subpiles, or to collapse piles into a smaller number. This permits a more refined scoring.

When only one level of sorting is used, the number of piles should range between three and eight. For the one-level case, the result of scoring for any one subject is a matrix A, in which a_{ij} is 0 if items i and j were not in the same pile and 1 if they were. When piles are further subdivided or collapsed, then a_{ij} is 1 if the two items were together only at the grossest level of the sort, 2 if they were together at the next more refined level, etc. Scaling and clustering techniques are used to combine the sorts of a group of subjects into a single picture of their cognitive structure. It is possible to assign a subject a score by measuring the departure of fit of his sorts from the prototypic expert-scaling solution.

Sorting tasks may be particularly sensitive to the restriction of range problems discussed above. That is, while experts and novices show strikingly different sorting solutions, we have not found very striking differences between higher and lower performers within a training cohort. Such differences as have been found seem to involve very small numbers of items that have specific ambiguities of nomenclature that only the better performers are sensitive to.

Characteristically, novices put things together in a sorting task on the basis of their superficial characteristics, while experts sort more on the basis of deeper meaning, especially meaning relevant to the kind of mental models or schemas that drive expert performance. For example, novice mechanics seem to treat parts with the same terms in their names as belonging together, while experts group more on the basis of the functional systems of which the objects being sorted may be parts. This difference is domain specific. That is, experts are not generally less superficial in their general world knowledge—only in the domain of their expertise.

When the technique is used to examine differences between people at the same level of formal training who have different levels of actual competence.

it affords an opportunity to discover which specific aspects of deeper understanding are not being picked up by the less able learner. This in turn can inform the design of improved instructional procedures.

Realistic Troubleshooting Tasks

In job domains that involve substantial amounts of diagnosis or other problem solving, some of the most revealing tasks used in cognitive task analyses are those that provide controlled opportunities for the subjects to actually do the difficult parts of their jobs. We have only begun to work on this approach, but a few possibilities already present themselves, particularly with respect to metacognitive skills of problem solving. To give a sense of our work, we trace the history of our efforts to analyze the performance of electronics technicians when they attempt to troubleshoot complex electronic circuitry. The complex cases are of particular interest because they are the ones where metacognitive skills are needed to organize processes that, in simple cases, might automatically lead to problem solution.

In our first attack on this problem, Drew Gitomer, at the time a graduate student at the Learning Research and Development Center, developed a troubleshooting task and simply collected protocols of subjects attempting to solve our problem. He then examined the protocols and attempted to count a variety of activities that seemed relevant to meta-cognitive as well as domain-specific skills. While the results, published in his thesis (Gitomer, 1984), were of great interest, we wanted to move toward a testing approach that was less dependent on the skills and training of a skilled cognitive psychologist. That, after all, is one aspect of what test development is largely about—rendering explicit the procedures that insightful researchers first apply in their laboratories to study learning and thinking.

Our breakthrough came not so much from deep cognitive thinking but rather from our interactions with an electronics expert who had extensive experience watching novice troubleshooting performances. He pointed out that it was not a big chore to specify all of the steps that an expert would take as well as all of the steps that any novice was at all likely to take in solving even very complex troubleshooting problems. That is, even when the task was to find the source of a failure in a test station that contained perhaps 40 feet[3] of printed circuit boards, cables, and connectors, various specific aspects of the job situation constrained the task sufficiently that the effective problem space could be mapped out.

[3] We are grateful to Gary Eggan for his many insights in this work.

This then created the possibility that we could specify in advance a set of probe questions that would get us the information we wanted about subjects' planning and other meta-cognitive activity in the troubleshooting task. In this most complex troubleshooting task, there are perhaps 55 to 60 different nodes in the problem space, and we have specific meta-cognitive probe questions for perhaps 45. Figure 1 provides an example of a small piece of the problem space and the questions we have developed for it.

An examination of the questions in the figure reveals that some are aimed at very specific knowledge (e.g., "How do you do this?"), while others help elaborate the subject's plan for troubleshooting (consider "Why would you do this?" or "What do you plan to do next?"). Combined with information about the order in which the subject worked in different parts of the problem space, this probe information permits reconstruction of the subject's plan for finding the fault in the circuit and even provides some information about the points along the way at which different aspects of the planning occurred.

Some of the scoring criteria that can be applied to the results of this blend of protocol analysis and structured interview are the following:[4]

(1) Did the subject find the fault? How close did he come?

(2) What kind of methods did the subject use? It may be possible to compare histograms of expert and novice distribution over the effective problem space to arrive at specific criteria for these decisions. For example, if there are points in the problem space that are reached only by experts, then getting to those points indicates expertise. Similarly, if there are sequences of steps that are present only in experts, the existence of such a sequence in the troubleshooting protocol of a person being tested might be counted positively.

(3) Was the subject explicitly planning and generating hypotheses about the nature of the problem? Did comments by the subject indicate that they had specific goals when they carried out sets of actions?

(4) Did subjects understand what they were doing? Did they have an answer to the question "Why did you do that?"

(5) Did the subject use available methods to constrain the problem space? Do planning and understanding components serve to help the subject constrain the search?

(6) Can the subject use available tools and printed aids? Can he use a schematic?

(7) How much information had to be given to the subject to enable him to continue with the problem? How long did it take the subject to complete subsections of the problem space?

[4] Debra Logan generated the first version of these criteria.

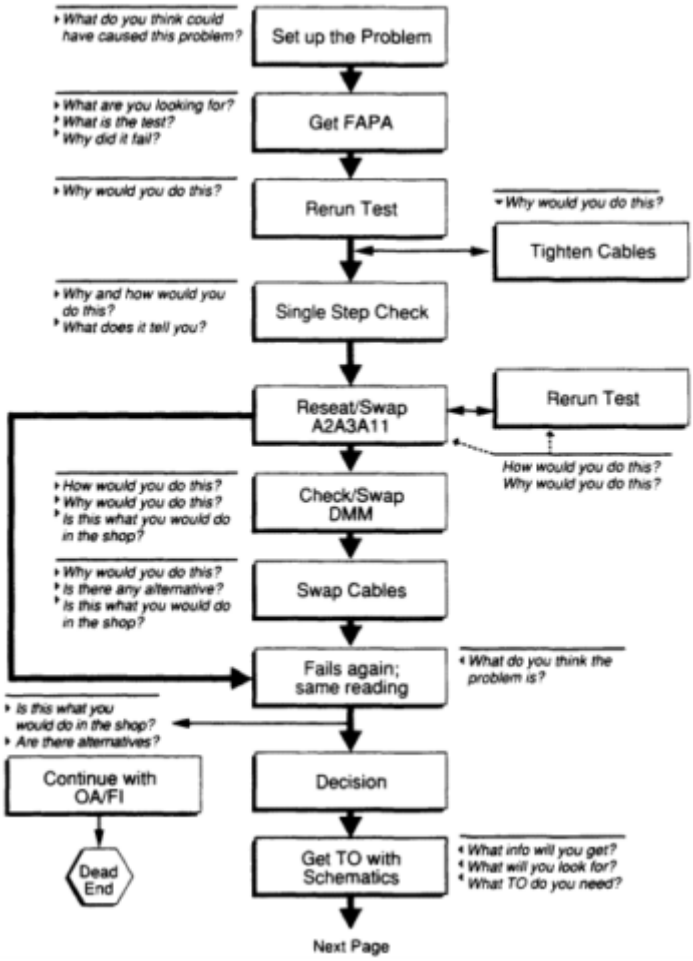


Figure 1 Problem space map to guide probed protocol gathering.

We are currently working on techniques for rating and scoring the responses to these individual components of the troubleshooting task and assessing how well they are integrated.

Connection Specification Tasks

A critical general skill for problem solving is the ability to break down a complex problem into smaller components and then tackle each component in turn. Unfortunately, there are sometimes interactions between components that preclude dealing with each one separately. For example, troubleshooting a device with five components by troubleshooting each component in turn will only work if the problem does not involve interconnections between components. If the problem is that the connector that joins two components is bad, the componential analysis approach based only on those components will fail. Knowledge of interactions between parts of a system, parts of a procedure, or parts of a problem is a critical component of expertise. In the case of system knowledge, this information can be extracted very directly: give the subject a sheet with all the components shown on it and ask them to show the interconnections among the components.

There are two ways such a problem can be presented. In one approach, the items are randomly distributed in such a way that no structure information is conveyed by the test form. Ideally, they should be at random points on the edge of a circle. Figure 2 shows such a display. Often, however, there is a superficial level of organization that all subjects are likely to share. In such cases, this organization might be given to subjects to minimize the extent to which searching the test form becomes part of the problem. An example is shown in Figure 3.

What-How-Why Tasks

Another task we have found very useful is one in which several basic kinds of knowledge about circuit components, tools, or other important job artifacts are measured. This task comes closest to overlapping traditional item types. Specifically, we can look at "what" knowledge, the ability to identify an object, to tell what it is. We can also look at "why" knowledge—what the object is used for. In addition, we can look at "how" knowledge, how it works. We have been quite successful asking such questions very directly and noting and recording the answers. This is preferable to the creation of multiple-choice items, which would also be possible, since sometimes the most interesting data for assessing level of competence is not right versus wrong but rather the terms in which a definition or specification of function is couched.

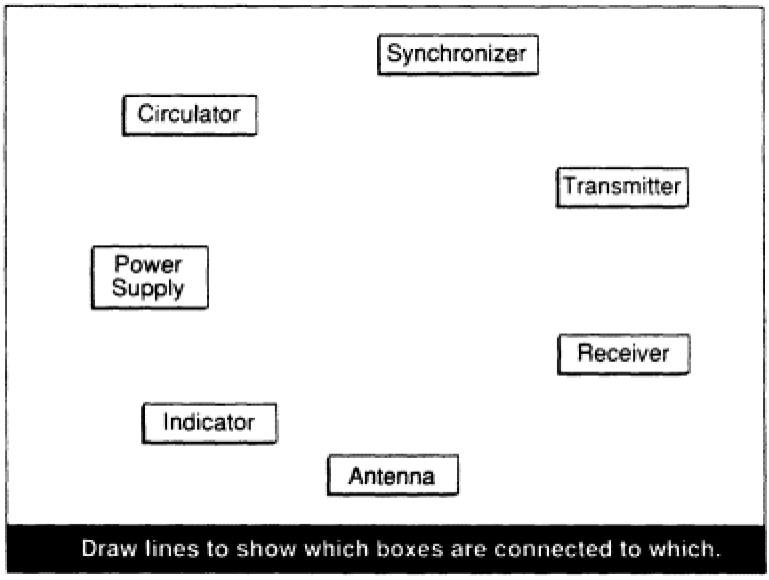


Figure 2 Randomly arranged connections test form.

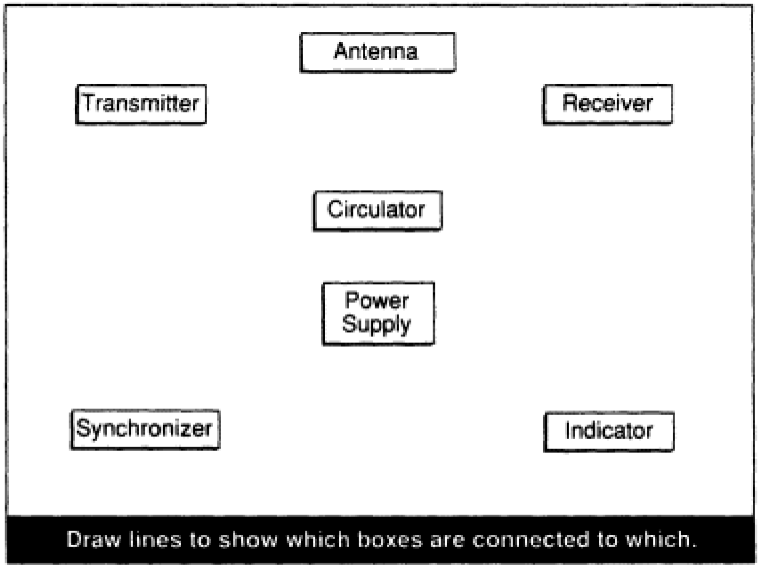


Figure 3 Systematically arranged connections test form.

AREAS OF THE MILITARY WHERE COGNITIVE TECHNIQUES HAVE PROMISE

Given the present high cost of cognitive task analyses and the need to use highly trained personnel because cognitive task analysis methods are still incompletely specified and validated,[5] such analyses should be restricted to situations in which a major investment of effort is appropriate. There are certain situations in which task analyses are likely to be productive given these current limitations. These are areas in which rational task analysis has not been able to supply adequate direction to those who design training systems because of the specific lack of an understanding of job difficulties that arise from the nature and limitations of human cognitive function.

In commenting on the specific situations in which cognitive approaches can be useful, we address three related topics. First is the work environment as it has been impacted by the hardware of modern technology. Next, we consider the interface between the worker and the work environment and examine various cultural influences on that union. Third, we consider some of the questions that decision makers in military systems must confront as they make selection, training, promotion, and job design decisions for complex tasks.

In conceiving of the sciences of the artificial, Simon (1981) characterized human performance (and learning) as moving across environments of varying complexity in pursuit of particular goals. With this conception, intelligent performance can mean, among other things, simplifying (and thus mastering) one's environment. Military work environments have grown steadily in complexity in recent decades as weapon systems, maintenance equipment, and other hardware used in the business of national defense have proliferated.

Interacting with complex machines in one's job is now the rule for military workers; however, the nature of intelligent performance in those interactions is not well understood. As a result, measurement of worker performance is often misguided because of the vagueness surrounding what it means to be skilled in a complex technical domain. Formal institutional attempts at simplifying the new high technology work environments consist mainly of thicker instructional manuals and technical documents. This suggests that details about complex systems are important for competence building. However, when we carefully analyze skilled performers to learn how they actually do their work, what we find are not detailed memorial replicas of

[5] These methods have yet to be validated. While there is no reason to believe that existing cognitive task analysis methods will produce misleading results, there is, however, some uncertainty in predicting the extent to which a particular method applied to a particular situation will produce an analysis that is a clear improvement over traditional methods.

dense technical data, but rather streamlined mental representations, or models, of the workings of the systems about which all the words are written. As skilled workers learn what their duties demand of them, they economically and selectively construct and refine their domain knowledge and procedural skill. That is one way they simplify their jobs. This suggests that the determinants of competence are not always revealed by the surface characteristics of either the worker's performance or the environment in which that performance takes place. Complex job environments require deeper cognitive analyses that can ferret out the conceptions and thinking that lurk behind observable behaviors.

For the military sector, understanding the influence of the machine on work and the dimensions of intelligent human performance in work settings is very important. Such understanding is crucial to decisions about

(1) the kind of intellectual talent needed for particular jobs (selection),
(2) the ways jobs and work environments could be constituted to optimize the use of available talent (classification),
(3) the instruction needed to build requisite skills (training),
(4) the basis for promotion,
(5) the optimal form for job performance aids and technical documentation, and
(6) the ways technical skills may be transferred across occupations (reclassification and retraining).

The complexity of the military workplace increases the difficulty of assessment to support these decisions. Targets of assessment are elusive in these settings.

Problem solving performance presents an interesting example. In the complex conditions we have been alluding to, problem solving cannot be fully programmed in advance because of the imprecise nature and generally illstructured state of the problems (Simon, 1965). Even with the mounds of technical data that exist, it is impossible to prespecify every problem solving scenario. Thus the assessment dilemma, if one is to capture true problem solving skill, is to measure nonprogrammed decision making, that is, to capture the intuitive reasoning that characterizes this form of expertise.

Recent work of cognitive psychologists interested in expert problem solving has advanced our understanding of reasoning and envisioning processes. For example, Larkin et al. (1980) have shown the rich sets of schemata indexed by large numbers of patterns that underlie the quickness of mind and insightful views of expert intuitive problem solvers, those who can fill in a sketchy representation with just the right pieces of information. Performance measures directed at such networks of influential knowledge in military job experts would be quite informative as predictors of competent problem solving.

A complex work environment further complicates performance assessment because of the inherent variability in human information processing. In a complex environment where so much is to be apprehended, encoded, and represented in memory, individual differences in cognition assume considerable importance. The interactions between a worker's cognitive apparatus (including all-important prior knowledge) and the many features of the complex systems encountered in the workplace are considerable. They seem as resistant to prespecification as the problem solving scenarios just discussed.

Even when a group of people enters the workplace after an apparently uniform initial job training experience, each person brings his or her own set of conceptions about the domain just studied. If performance assessment is directed toward the measurement of individual skill for purposes of improving performance, i.e., if the goal is diagnosis to prescribe instruction, then individual differences in cognition are worth some attention. A cognitive analysis that examines complex human performance in depth may uncover uniformities as well as common misconceptions or bugs that affect the learning process. This can lead to better design and development of the kind of adaptive training that can significantly facilitate learning in complex domains.

There are sources of influence external to the workplace that interact with it to affect even further a worker's performance. These influences have implications for both cognitive analysis and performance assessment. First of all, there are the pressures of the military culture. The apprentice is pressed to learn the job quickly in order to become a contributing member of the work force as soon as possible. Typically, it is the cadre of apprentices who are relied on as the critical mass or core capability of a military operational unit. Simultaneously, however, the demand characteristics of a military unit can be quite severe, which is to say that putting planes in the air, for example, takes precedence over on-the-job training sessions. Learning the job quickly is thus frequently impeded because of the demand to get the work out at all costs. Opportunities to practice and refine skills are characteristically nonexistent, particularly for the worker of average skill or below.

A potential influence on learning and job competence comes as a consequence of the modern tactical philosophies currently favored by some of the Services. In the interest of dispersing weapon systems and maintenance teams for purposes of reducing concentrated resources as inviting targets, actions are being considered to make weapon system maintenance occupations less specialized. By assigning broader responsibilities to a given class of worker, fewer technicians, who would be transformed from specialists to generalists, would be required in field locations to perform maintenance functions. Such a policy would necessitate dramatic changes in instructional practice in order for broader domain knowledge and more flexible reasoning skills to be realistic targets of training.

Already training practices have been weakened under the weight of complex subject matter and formidable workplace machines. An argument that has sometimes prevailed is that smarter machines mean reduced cognitive loads on workers and that consequently less training is required. Of course, machines capable of automating certain workplace tasks, i.e., the relatively easy portions of the jobs, do not in reality appreciably reduce the cognitive workload. Rather, what the machines do is take responsibility for the lower order or programmed tasks, reducing the apprentice to a passive observer who is called into action only when nonprogrammed problem solving is required. In other words, the apprentice loses opportunities to learn by doing some of the routine workplace tasks but is expected to somehow acquire the ability to solve problems either when the machine breaks down or when the problem is beyond the machine's capabilities.

The complex machines also pose logistics problems to the training community who have had the formidable task of evaluating the increasingly complex workplaces of the military to determine instructional goals. Tough questions about the fidelity of the training place vis-à-vis the workplace have been vigorously debated. The training place typically has low priority for the expensive machines that populate the workplace, an unfortunate consequence of which is that during training, hands-on learning opportunities are often replaced by theoretical abstractions that cannot be tied to concrete experience.

All of this translates into the following kind of scenario for the typical apprentice: initial technical training is customarily patterned on an academic model of teaching complex subject matter. Students are told about a work domain instead of receiving practice in it. The academically-trained apprentice is met at the workplace by high expectations and by demand characteristics that simultaneously increase the pressure to learn and eliminate many learning opportunities. The implications for the interplay of performance assessment and cognitive analysis in this context of labored apprenticeship learning can be summarized in the following points:

(1) Inventive testing informed by cognitive analyses could conceivably begin to shift the emphasis in technical training away from academic models of learning facts to experiential models of learning procedures. Frederiksen (1984) has reported precedents in military training where changing the test meant instructional reform. The reason for present assessment being focused on declarative knowledge is a familiar one in psychology, namely, that what usually gets measured is that which is easy to measure (e.g., the formula for Ohm's Law versus facility in tracing signal flow). Results of cognitive analyses of procedural knowledge provide a rich basis for constructing items that do more than test recognition skill.

(2) Cognitive testing approaches are characterized by the methodology

employed in creating test items and not necessarily by the form of those items. We believe that cognitive theory now poses important issues to be considered in evaluating a particular approach to testing. Specifically, the approach we favor is one of identifying the critical mental models, conceptual knowledge, and specific mental procedures involved in competent performance and then asking whether a given test allows one to reliably assess the extent of those aspects of competence. This suggests that traditional paper-and-pencil formats may have to be supplemented by hands-on testing in order to be sure that procedural skills are well established, but it also suggests that even exhibiting competent performance on the job may not predict transfer capability nor the ability to work well with nonstandard problems or work conditions. Thus, cognitive approaches may require that direct demonstrations of competence on "fair" problems under safe and standard conditions be supplemented by computer-based testing that can simulate unsafe, expensive, and otherwise nonstandard problem solving contexts.

(3) Computer delivery of diagnostic items affords opportunities for testing environments to double as adaptive learning environments. Intelligent simulation environments are feasible as well where work instruments can be represented for learner exploration, manipulation, even simplification. This kind of microworld approach is an interesting way to cope with the absence of real systems in the training place. Likewise, the computer microworld can move to the workplace to introduce a constant source of on-the-job training and practice experiences.

(4) Finally, the prospect of broadening a worker's technical purview presents the dual challenge of uncovering knowledge and skill components that cut across existing specialized occupations and devising instruction that generates transfer. Both parts of the challenge entail performance assessment demands. Cognitive theory-based work following the expert-novice paradigm has amassed some evidence to suggest commonalities in expertise across domains such as physics, electricity, and radiology—e.g., deep versus surface structure in problem representation, knowledge in highly proceduralized form (Chi et al., 1981; Gentner and Gentner, 1983; Lesgold et al., 1988). Components of skill like these, that span multiple domains, represent logical foci for instruction and assessment where movement across domains is of interest.

REFERENCES

Anderson, J.R. 1976 *Language*, *Memory*, *and Thought*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

1982 Acquisition of cognitive skill. *Psychological Review* 89:369-406.

Card, S.K., T.P. Moran, and A. Newell 1983 The Psychology of Human-Computer Interaction. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Chi, M.T.H., P. Feltovich, and R. Glaser 1981 Categorization and representation of physics problems by experts and novices. *Cognitive Science* 5:121-152.

Chi, M.T.H., R. Glaser, and E. Rees 1982 Advances in the psychology of human intelligence. In R. Sternberg, ed., *Expertise in Problem Solving*, Vol. I. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Fitts, P.M. 1964 Perceptual-motor skill learning. In A.W. Melton, ed., *Categories of Human Learning*. New York: Academic Press.

Frederiksen, N. 1984 The real test bias: influences of testing on teaching and learning. *American Psychologist* 39(3):193-202.

Gentner, D., and D. Gentner 1983 Flowing waters or teeming crowds: mental models of electricity. In D. Gentner and A.L. Stevens, eds., *Mental Models*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Gitomer, D. 1984 A cognitive analysis of a complex troubleshooting task. Unpublished dissertation, University of Pittsburgh.

Kieras, D.E., and P.G. Poison 1982 An approach to the formal analysis of user complexity. Working Paper 2, Project on User Complexity of Devices and Systems. University of Arizona and University of Colorado.

Larkin, J.H., J. McDermott, D.P. Simon, and H.A. Simon 1980 Expert and novice performance in solving physics problems. *Science* 208:1335-1342.

Lesgold, A.M. 1984 Acquiring expertise. In J.R. Anderson and S.M. Kosslyn, eds., Tutorials in Learning and Memory: Essays in Honor of Gordon Bower. San Francisco: W.H. Freeman.

1986 Problem solving. In R.J. Sternberg and E.E. Smith, eds., *The Psychology of Human Thought*. Cambridge, Eng.: Cambridge University Press.

Lesgold, A.M., and C.A. Perfetti 1978 Interactive processes in reading comprehension. *Discourse Processes* 1:323-336.

Lesgold, A.M., L.B. Resnick, and K. Hammond 1985 Learning to read: a longitudinal study of word skill development in two curricula. Pages 107-138 in T.G. Waller and G.E. MacKinnon. eds., Reading Research: Advances in Theory and Practice, Vol. 4. New York: Academic Press.

Lesgold, A.M., H. Rubinson, P.J. Feltovich, R. Glaser, D. Klopfer, and Y. Wang 1988 Expertise in a complex skill: diagnosing X-ray pictures. In M.T.H. Chi, R. Glaser, and M. Farr, eds., *The Nature of Expertise*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Newell, A., and H.A. Simon 1972 *Human Problem Solving*. New York: Prentice-Hall.

Schneider, W. 1985 Training high-performance skills: fallacies and guidelines. *Human Factors* 27 (3):285-300.

Simon, H.A. 1965 *The Shape of Automation*. New York: Harper and Row.

1981 *The Sciences of the Artificial*. Cambridge, Mass.: MIT Press.

Work Samples as Measures of Performance

Frederick D. Smith

INTRODUCTION

The ability to predict future performance of incumbent employees or candidates for employment has been one of the main contributions of psychologists to industry. Traditionally, this process has involved interviews, psychological testing, and biographical information as predictors of on-the-job performance. The predictive validities of these methods, while often good enough to satisfy legal requirements, are not as high as could be expected given the long history of research and development in testing and interviewing.

One method that appears to overcome some of the problems of these traditional selection techniques is the work sample. Work samples measure job skills by requiring an individual to demonstrate competency in a situation parallel to that at work, under realistic and standardized conditions. Their primary purpose is to evaluate what one can do rather than what one knows (Cascio and Phillips, 1979).

This paper is a review of the work sample literature, with particular attention paid to the theory underlying work sample testing, the use of work samples as criterion measures, the adverse impact of work samples, and measurement issues associated with such tests. In order to understand the work sample in its nontraditional use as a criterion measure, however, it was felt that some of the research employing work sample tests as predictors

would be illuminating. For this reason, a section of the paper is devoted to some representative studies using work samples as predictors.

A work sample, whether used as a predictor or a criterion, is constructed to allow a measure of performance on a structured task that is directly reflective of the types of behaviors required in the job situation. Consequently, regardless of its use in the literature, either as a predictor of future performance or as a measure of present ability, the work sample by its very nature can be considered a criterion measure.

In a similar vein, criterion work samples that appear in the literature are most often used to gauge success in training. To the extent that success in training is felt to be a predictor of eventual performance on the actual job, the work samples used in this manner serve the dual purpose of measuring current performance and predicting future success.

For both of these reasons, the author feels that the predictor/criterion distinction is superficial in the case of work sample testing, and that including certain predictive work samples in this review will be beneficial for an understanding of these types of tests.

Theoretical Bases for Work Sample Testing

Theoretically, work samples should possess high validity since the test itself is a subset or a sample of the criterion domain. Wernimont and Campbell (1968) have suggested that performance prediction based on work samples would be fruitful. They propose a behavioral consistency model founded on the tenet that the best predictor of future performance is past performance. In applying their approach, Wernimont and Campbell recommend searching an applicant's work experience for specific examples of required job behaviors, and if these do not exist, using a work sample or simulation. In proposing this view, Wernimont and Campbell (1968:372) hope to overcome what they see as "the unfortunate marriage" of the "classic validity model with the use of tests as signs, or indicators of predispositions to behave in certain ways, rather than as samples of the characteristic behavior of individuals."

Similarly, Asher (1972) and Asher and Sciarrino (1974) suggest that predictive power is enhanced when there is a point-to-point correspondence between the predictor and the criterion space. For this reason, tests of a single dimension are less powerful predictors than more complex tests such as work samples, which are designed to be miniature replicas of the criterion task. While the validities reviewed later in this paper seem to bear out the predictive power of work samples, Asher and Sciarrino themselves mention several other possible explanations. One is an interaction hypothesis. A complex task may elicit an interaction among aptitudes rather than a simple additive effect, so that the criterion will be poorly predicted based on an additive model using measures of single aptitudes or traits. If a prediction

model is built by combining aptitude test scores additively, the model may in fact overlook aptitudes that interact. A work sample, by its design, allows these interactions to occur naturally, and would therefore be expected to predict future performance to a greater degree than a series of individual tests.

Another possible explanation for the high validities of work samples is that of work methods. The work sample may elicit realistic work habits that individuals use to solve specific problems, and these work methods may account for a greater amount of individual differences than combinations of basic motor or verbal abilities tapped by paper-and-pencil tests. By having individuals perform actual work-related behaviors, they are able to demonstrate ability more specific to the job itself rather than some generalized aptitude. The point of a work sample is, after all, to reduce the inferential leap that must be made between performance in a standardized testing situation (be it motor or written or verbal) and actual job performance. There is less of a leap needed between behavior in a work sample and behavior in the actual job situation than between performance or problem solving on a paper-and-pencil test and actual job behavior.

A final point made by Asher and Sciarrino (1974), and one somewhat related to that of work methods, is that experience may play a role in the high validities found in work sample testing. The work sample may be measuring prior experience that has transferred to the criterion task, and may therefore be identifying people with more experience but not necessarily more aptitude.

In addition to these comments regarding the point-to-point theory, Gordon and Kleinman (1976) suggest that the face validity of work samples may influence motivation among testees, which is related to the interest in and motivation for a particular job. Therefore, a less identifiable set of elements than point-to-point correspondence between predictor and criterion may be responsible for high validities reported. The Gordon and Kleinman study is examined in further detail below.

Many of these same points can be made for work samples as criterion measures. If the purpose of a work sample is to obtain an accurate measure of current job performance, then the work sample must accurately reflect job behaviors critical for success. Rather than rating employees on a number of gross performance dimensions, the work sample allows appraisal on a specific, standardized, and possibly complex task. For example, an appraisal on the independent dimension of troubleshooting is a poor measure compared to appraisal on a complex task that requires troubleshooting for successful completion. In traditional performance appraisal, correlation between dimensions is halo and is often thought of as error; in a work sample task, the natural and perhaps critical correlation between several abilities or dimensions can occur as part of the testing process.

Similarly, work methods can be accurately reflected in a work sample criterion. The techniques that an employee actually uses in the job situation can be observed and measured, rather than some generalized performance dimension of mechanical ability or problem solving.

Experience would also be expected to play a role in work sample criterion measures. This would be especially true if the task required a substantial amount of training or job specific knowledge. More experienced employees should be expected to perform better than those less experienced. For this reason, care must be taken that a work sample meant to measure job performance note the ability level of testees. A single work sample may not be appropriate for all employees, or different performance standards may be needed.

The theoretical underpinnings of work samples just reviewed imply that they are superior measures of performance. Empirical results seem to support this contention. What follows is a summary of several reviews of work samples as predictors, and then a more detailed look at the work sample as a method of measuring performance.

Previous Reviews of Work Sample Testing

There have been a number of reviews containing validity data on work sample tests. Asher and Sciarrino (1974) classified over 60 work sample studies into either motor or verbal tasks. A work sample was considered motor if it involved the physical manipulation of things, such as operating a sewing machine, tracing a complex electrical circuit, or repairing a gear box. Verbal work sample tests involved language-oriented or people-oriented tasks. These included tests of common facts in law for students, in-basket exercises, role plays for making telephone contacts with customers, and skill in writing business letters. The criteria in the studies reviewed were either job proficiency or success in training, and criterion measures were generally supervisor ratings, output (number of items produced), completion of training, or grade in training. For some of the verbal work samples, criteria also included salary, promotions, job level, sales, or number of leadership offices held.

As can be seen in Table 1, with job proficiency as the criterion, motor work samples were second only to biographical information in terms of predictive validity. Forty-three percent of the motor work samples reviewed had validity coefficients greater than .50, and 70 percent of the motor work samples had validity coefficients exceeding .40. Verbal work samples fared less well when job proficiency was the criterion, with only 21 percent of the validity coefficients exceeding .50, and 41 percent exceeding .40. However, with success in training as the criterion measure, verbal work samples were superior to motor work samples. Thirty-nine percent of the verbal work

TABLE 1 Proportions of Work Sample Validity Coefficients Exceeding Particular Levels

	Criterion					
	Job Proficiency			Success in Training		
Predictor	$r \geq .50$	$r \geq .40$	$r \geq .30$	$r \geq .50$	$r \geq .40$	$r \geq .30$
Motor work samples	.43	.70	.78	.29	.47	.79
Verbal work samples	.21	.41	.60	.39	.65	.81

SOURCE: Asher and Sciarrino (1974).

samples had validity coefficients in excess of .50, while only 29 percent of the motor work samples' validities were larger than .50. Sixty-five percent of the verbal work samples reviewed had validity coefficients greater than .40, while only 47 percent of the motor work samples' validities were greater than .40. Asher and Sciarrino (1974) conclude from their review that work samples fare well when compared to other predictors, being second only to biographical information in terms of predictive power.

A more recent review of the work sample was performed by Robertson and Kandola (1982). They divided 60 work sample tests into four categories: psychomotor, individual situational decision making, job-related information, and group discussions/decision making. The psychomotor category corresponds to Asher and Sciarrino's motor tests, while the other three could be considered more verbal. Criteria included job performance, job progress, and training. It was found that psychomotor tests had a median validity of .39 (78 coefficients), job-related information tests a median of .40 (27 coefficients), situational decision making a median of .28 (53 coefficients), and group discussion tests a median of .34 (27 coefficients).

Table 2 was constructed from data appearing in the Robertson and Kandola (1982) review, and allows some comparison to the Asher and Sciarrino results. Table 2 shows the proportions of validity coefficients exceeding particular levels for the four types of work sample predictors with job performance and training performance as the criteria. (The only predictive validities reported for the criterion training were for psychomotor tests.) A pattern of results similar to those of Asher and Sciarrino can be found. When job performance is the criterion, psychomotor work samples outperform work samples that are more verbal. Comparing psychomotor validities across the two criteria, it can be seen that they predict less well for training than for job performance.

An important difference between the earlier reviews by Asher and Sciarrino and Robertson and Kandola, which is of particular relevance to this paper, is the use of a work sample as a criterion measure. The Robertson and

TABLE 2 Proportions of Work Sample Validity Coefficients Exceeding Particular Levels

	Criterion					
	Job Proficiency			Training		
Predictor	$r \geq .50$	$r \geq .40$	$r \geq .30$	$r \geq .50$	$r \geq .40$	$r \geq .30$
Psychomotor	.31	.69	.88	.16	.50	.75
Job-related information	.09	.27	.36			
Situational decision making	.08	.27	.50			
Group discussion	.30	.40	.80			

SOURCE: Based on Robertson and Kandola (1982).

Kandola review is not complete, since they only include studies in which a work sample was used both as the predictor and the criterion. As a criterion, the work samples reviewed were usually similar to but longer than the predictor work sample. Robertson and Kandola report a median validity of .49 between psychomotor predictors and work samples consisting of psychomotor tasks (based on 10 validity coefficients). A median validity of .75 (based on 7 validity coefficients) was found between situational decision making predictors and a work sample criterion of performance in an assessment center.

While these validities are impressive, Robertson and Kandola caution that the idea of increasing the similarity between predictor and criterion (as per point-to-point theory, for instance) may have been pushed beyond a reasonable limit. These correlations can be interpreted as measures of reliability rather than validity. By comparing one job-related test with another job-related test, the relationship between the two tests may be discovered, but inferences of how this relationship relates to job performance will still have to be made. It is precisely this inference that work sample testing is supposed to reduce. Robertson and Kandola (1982) caution that researchers should not attempt to increase validity by simply developing criteria that are likely to relate closely to the predictor. Rather, care should be taken that the criteria themselves are job performance measures.

Meta-analytic Reviews of Validity Studies Involving Work Samples

Two recent articles reviewed validities of predictors of job performance. Schmitt et al. (1984) performed a number of meta-analysis on validity studies published between 1964 and 1982. Their analyses revealed that work samples, assessment centers, and superior/peer evaluations yielded validities superior to general mental ability tests or special aptitude tests. When

work samples are used as predictors, the average validity coefficient was .378 (based on 18 coefficients); when work samples were the criterion, the average validity was .401 (24 coefficients).

Hunter and Hunter (1984), using meta-analytic techniques, found that for entry level jobs for which training will occur after hiring, combined cognitive ability and psychomotor ability test scores had a mean correlation of .53 with performance on the entry level job (425 validity coefficients), while a job tryout had a mean correlation of .44 with the same criterion (20 coefficients). For selection on the basis of current job performance, the work sample was slightly better than the ability composite, with average validity coefficients of .54 and .53, respectively. In all these cases, the work sample served as a predictor. Hunter and Hunter also report a meta-analysis involving studies in which work sample performance was used as a criterion and a job knowledge test was the predictor. A mean validity of .78 was obtained (based on 11 coefficients). The authors note, however, that job knowledge tests can be used for prediction only if the examinees are already trained for the job.

Hunter and Hunter's (1984) results differ somewhat from those of Schmitt et al. (1984), and this could be due to the fact that Hunter and Hunter (1984) include a large portion of unpublished data in their study, while Schmitt et al. drew data from published studies in the *Journal of Applied Psychology* and *Personnel Psychology*. In any case, again it appears that work samples have validities comparable to, and in many cases, superior to other predictors. While their use as criteria has been more limited, these two meta-analytic reviews do report rather impressive average validity coefficients for work samples as criteria.

WORK SAMPLES AS PREDICTORS

Work samples have traditionally been used as predictors of future job performance. This section is intended as a brief review of these types of studies.

Campion (1972) developed a work sample test for maintenance mechanics in a food processing company. The work sample consisted of four tasks, each broken down into the number of steps required to complete it. The tasks were: installing pulleys and belts, disassembling and repairing a gearbox, installing and aligning a motor, and pressing a bushing into a sprocket and reaming it to fit a shaft. In addition to these work sample tasks, each mechanic was given several paper-and-pencil tests: the Test of Mechanical Comprehension, Form AA; the Wonderlic Personnel Test, Form D; and the Short Employment Tests. Criterion measures were supervisor evaluations of three factors: use of tools, accuracy of work, and overall mechanical ability. It was found that performance on the work sample was significantly and positively correlated with supervisor evaluations of work performance on

all three criteria, but that none of the validity coefficients for the paper-and-pencil tests was statistically significant.

Gordon and Kleinman (1976) also compared a work sample test to a paper-and-pencil test, with the criterion being training scores. Three classes of recruits in a police training academy were given a work sample test including firearms and defense tactics (motor tests), and a written work sample addressing the relationship of the police department to other civic agencies, department rules and regulations, and an introduction to law enforcement. A general intelligence test, the Otis-Lennon Mental Ability Test: Form J, was also administered. For all three classes of recruits, the work sample scores predicted overall training scores, while the intelligence test was significantly correlated with the criterion for only one class. As mentioned previously, Gordon and Kleinman suggest that the face validity of work sample tests may influence the motivation of testees, and this is also related to the interest in the job.

A study that failed to find any correlation between the work sample and a performance criterion was reported by Inskeep (1971). Three work sample tasks used to select and place sewing machine operators were examined using a concurrent validation design. The tests were developed to reflect actual shirt-making operations. The clipboard test uses a table with a sliding center board on which are mounted a number of metal clips. When the center board is moved into proper position, a clip may be opened by a foot pedal linked to the table top. The subject is provided with two piles of cloth rectangles. The subject must pick up a rectangle from each pile, align them, and place them in a clip. Then he or she slides the center board to align the next clip and repeats the procedure until all clips are filled. Performance score is the total time to fill all clips.

The needle board is the second work sample task. Ten spindles of thread and a metal crossbar are mounted on a table. In the crossbar are 10 needle holes corresponding to the 10 spindles. The subject is required to pass the thread through both a vertical and a horizontal needle hole. The score for the test is total time required to complete all 10 threadings.

The final test is called the hurdles. This involves using a standard production sewing machine geared down to a lower operating speed. The subject must sew along a specified pattern and complete a certain number of stitches. Test score is the number of seconds required to complete the sewing exercise.

Inskeep (1971) used a performance criterion of piece-rate earnings. It was found that the correlation between the clipboard test and earnings was -.02, between the needle board test and earnings was-.06, and between the hurdles and earnings was-.08, all nonsignificant. The Inskeep findings are somewhat surprising in that the work sample tasks are almost identical to some of the actual job behaviors required of incumbent sewing machine

operators. This may reflect a problem with the performance criterion of piece-rate earnings, although Inskeep did not offer possible reasons for the negative findings.

A work sample test in the form of a minicourse for telephone switching repairmen was examined by Reilly and Manese (1979). The minicourse was a short (about 40 to 60 hours) training program designed to be a content valid sample of a 6-month electronic switching system (ESS) course. Predictors were total time to complete the minicourse and test performance based on seven self-paced lessons on electronic switching system fundamentals, plus the score on an ESS minicourse summary test. The criteria were total time to complete the full electronic switching system course, which consisted of two separate self-paced courses, one containing four modules, and the other containing five modules. It was found that minicourse test scores were significantly and negatively correlated with time to complete the full course, and that time to complete the short electronic switching system course was significantly correlated with time to complete the full course. Reilly and Manese (1979) comment that since the average cost per trainee for the long electronic switching system course is $25,000, the cost benefit of a valid selection procedure can be substantial.

Assessing Trainability Using Work Samples

It appears, then, that a work sample can be a valid means of assessing trainability of job candidates. Robertson and Downs (1979) distinguish between work sample tests and trainability tests: trainability tests include a structured and controlled period of learning and are used to select personnel for training rather than to choose people who are already competent. The procedure usually involves three steps:

(1) Using a standardized form of instruction and demonstration, the instructor teaches the applicant the task, during which time the applicant is free to ask questions.
(2) The applicant performs the task unaided.
(3) The instructor records the applicant's performance and also makes a rating of the applicant's likely performance in training.

They review 16 studies, in which 24 validities are reported. The criterion in most is training success. Of the 24 correlations, 20 are significant, with coefficients in excess of .50 found in 10 cases. Robertson and Downs (1979) conclude that trainability tests display high content and face validity and allow the applicant to get a clear understanding of the job in question (a realistic job preview, in a sense), but that they are very job-specific and need to be redesigned and revalidated as jobs change, as well as being expensive to administer.

Robertson and Mindel (1980) examined the correlations between trainability tests and performance after 3 weeks of training for six craft trades: bricklaying, capstan, carpentry, milling, welding, and center lathe. They found that for three of the crafts, scores on a trainability test correlated significantly with training performance. Robertson and Mindel caution that the lack of predictive validity for some tests illustrates that although there is a generalized procedure for designing and administering the tests, each must be validated individually.

It would seem to be a short methodological step from using a work sample as a predictor of training ability to using a work sample as a measure of training success and also as a job performance criterion. The next section begins by examining a study that used a work sample both during training and to evaluate on-the-job performance. This will be followed by a review of studies in which work samples are primarily the criterion measures.

WORK SAMPLES AS CRITERIA

Relatively few studies have employed work samples as criterion measures, and those that do generally measure training achievement. Work samples used as criteria are useful because they provide a standardized testing situation in which to evaluate employees, and would seem to lend themselves well to jobs that are highly structured or jobs for which a core of representative behaviors could be identified and developed into a work sample.

One particular study nicely bridges the gap between work samples as predictors and the use of work samples as criteria. Siegel and Bergman (1975) developed what they called the miniature job training and evaluation approach. This is similar to trainability testing in that the examinee is trained, through demonstration and practice, to perform a particular task. The examinee is then scored on how well he or she performs what was taught with regard to following proper procedures, safety, and care and use of tools. The approach is based on demonstration of the ability to learn parts of the job as predictive of total job success.

Subjects in the Siegel and Bergman (1975) study were low aptitude U.S. Navy recruits who had failed a standardized paper-and-pencil test for admission into machinist's mate school. The paper-and-pencil test had three parts: a general classification test, an arithmetic test, and a mechanical test.

In developing their training program, the authors identified six behaviors as most representative of those performed by a journeyman machinist's mate: tool identification and use, gasket cutting, meter reading, trouble shooting, equipment operation, and assembly. Training sessions of 15 to 30 minutes were built around these behaviors. Once training was completed, each subject was tested on the amount learned during the training phase. The test was a procedural review of what was taught during the training session.

Following the completion of training, subjects were assigned to the fleet for duty.

Criterion measures were taken 9 and 18 months following completion of the training program. The criterion tasks were developed based on the opinions of experienced Navy chief machinist's mates. These reflected a diversified sample of the range of behaviors involved in the job of journeyman machinist's mate, and included the following: standing messenger watch, breaking and making a flange, packing a valve, demonstrating procedures in common malfunction and in emergency situations, knowledge of use and names of common equipment and tools, manifesting general alertness and common sense in the work situation, and adequacy of technical job knowledge. These criteria were administered individually to each of the subjects at the 9- and 18-month follow-ups. At the first criterion follow-up, 54 of the original sample of 99 subjects were available for testing, and 34 of the original 99 subjects were available for the second criterion follow-up. Siegel and Bergman do not say whether any of the subjects were the same from the 9and 18-month follow-ups.

In order to compare the Navy paper-and-pencil predictors with their work sample predictors, Siegel and Bergman created a composite criterion score. The three work sample predictors with the highest zero correlation with the composite criterion (gasket cutting, trouble shooting, and assembly) were then used to determine the multiple correlation with each of the criterion tests. Siegel and Bergman reasoned that since only three predictors were used in the standard Navy selection technique, they would employ only three predictors.

For the 9-month criterion test, significant multiple correlations were found between the work sample and the standing messenger watch, knowledge of equipment and tools, and alertness and common sense criteria. The Navy predictors were correlated with knowledge of equipment and tools and with alertness and common sense. Disregarding significance levels, five of the seven performance criteria were predicted better by the training work samples than by the Navy tests, and Siegel and Bergman find this directional difference significant using a sign test.

At 18 months, directly opposite results were found. There were significant multiple correlations between the Navy tests and all but one of the criteria (alertness and common sense), while none of the criteria were predicted by the work sample training scores.

Siegel and Bergman conclude that the miniature job training and evaluation concept possesses merit for predicting performance of low aptitude applicants. They suggest that the lower predictive power of the work sample scores over time may be due to basing predictions on specific training scores rather than general abilities. While the job training work samples may be adequate for predicting success at initial job entry, over time

continued success depends on generalized verbal and conceptual factors. In other words, the work sample training was very specific and applied in nature, while the Navy predictors measure a more generalized ability that remains stable over time. While this is a possible explanation for the results, it certainly does not help their case for using the miniature job training approach in place of traditional Navy selection tests. In effect they are saying that while their type of work samples are useful for getting low aptitude candidates started in the craft positions, these individuals are still lacking in some important basic aptitudes that make continued job success problematic.

In addition to this conceptual problem, Cohen and Penner (1976) identify several methodological problems with the Siegel and Bergman study, among them the improper use of a sign test, the lack of cross validation, and the fact that Siegel and Bergman (1975) performed a discriminant analysis and a validation study on the same sample. Siegel (1983), in a follow-up study employing larger samples and a greater number of Navy job specialties, again found a modest number of significant validities between miniature job training predictors and performance at 9 and 18 months. The criterion measure in this study, however, was not a work sample but commanding officer ratings of a subject's performance on technical aspects of his or her work. The ratings were on a 7-point scale ranging from "very poorly" to "very well." Siegel again concludes that the miniature job training approach shows good predictive validity for the 9-month period, less so for the 18-month period, and that the approach has merit compared to traditional paper-and-pencil testing.

Physical Ability as a Predictor of Work Sample Performance

A work sample was developed to validate selectors for filling steelworking positions on the basis of physical ability (Arnold et al., 1982). Work samples for entry level positions in the general labor pool were developed based on job analyses and interviews with managers and incumbent laborers. Some of the work sample activities were shoveling slag, lifting and moving 75-pound bags, carrying jackhammers, and wheelbarrowing. An abstracted work sample was then developed that tapped the general physical abilities required to perform the tasks in the work sample. This abstracted work sample drew on the work of Fleishman (1964) and included static strengths, dynamic strengths, balance, and flexibility. These abilities were the predictors in the study, and were measured by arm, leg, and back dynamometers, balance beam, leg lifts, push ups, squat thrusts, pull ups, and a step test. It was found that the correlations between the abstract work sample items tapping strength and the work sample performance measures were consistently high across three worksites: 82 percent of the correlations were above

.40. The arm dynamometer was found to have particularly high correlations with work sample performance: average correlation across three work sites with a composite work sample measure was .84. Using multiple regression analyses, Arnold et al. (1982) conclude that the arm dynamometer measure alone is sufficient for selection purposes. They also conclude that using a common regression line would have only a slight bias toward men. Thus, a work sample was successfully used to validate the strength test, and Arnold et al. report that using the arm dynamometer as a selection device could potentially save the company over $9 million a year (using the Hunter et al., 1979, utility techniques).

Another study that used physical ability tests as predictors of work sample performance was by Reilly et al. (1979). In testing telephone company craft jobs, they found that dynamic arm strength and reaction time correlated .34 and-.33, respectively, with an overall work sample performance score. The work sample included pole testing, climbing stepped poles, placing ladders on a cable, placing ladders on a building, climbing unstepped poles, and climbing unstepped poles and removing a drop wire. A common regression line could be used for both males and females.

Paper-and-Pencil Tests as Predictors of Work Sample Performance

Frank and Wilcox (1978), in a study of 22 firemen, used a 6-hour work sample of firefighting skills as a criterion to cross-validate the Strong-Campbell Interest Inventory's moderating effect on the Raven's Progressive Matrices (short form) intelligence test. The work sample covered seven major areas: handling hose lines, ladders, ropework, ventilation procedures, first aid, small tool knowledge, and use of oxygen masks. However, due to lack of variability, the last four tasks were eliminated from the test battery. A single fire captain, who had no previous contact with the trainees, evaluated individual segments of each task as either being performed right or wrong, and then assigned an overall score to each area. It was found that for subjects above and below the median on the Strong-Campbell, the correlation of the Progressive Matrices with the criterion was significantly different. No racial bias was found in either the moderator or the predictor.

In a study of 211 minority and 219 nonminority telephone company repairmen and installers, Grant and Bray (1970) validated five aptitude tests against proficiency measures obtained from a learning assessment program. The learning assessment program is organized into seven levels of training, in ascending order of difficulty, and includes basic electricity, basic telephone, Bell System practices, station circuits, advanced circuits, and trouble location. The learning assessment program is programmed, and a trainee continues through the seven levels until he or she fails to meet the requirements of a particular section. Grant and Bray found that all of the aptitude

tests were predictive of success in the learning assessment program, and that correlation coefficients between the minority and nonminority samples were comparable. In addition, regression equations for the two samples were compared, and it was found that the slopes were almost identical but that the intercepts differed slightly.

A series of studies by Gael and others (Gael and Grant, 1972; Gael et al., 1975a, 1975b) used work sample criteria to validate employment tests for several telephone company occupations. In the 1972 study, minority and nonminority service representatives completed a general learning ability test (the Bell System Qualification Test I), five clerical aptitude tests (spelling, number comparison, arithmetic, number transcription, and filing), and a role-play interview modeled after actual service representative contacts with customers. There were two general criteria. A paper-and-pencil achievement test was used to measure comprehension and retention of company policies and job procedures and practices. The second criterion was a work sample composed of typical calls that a service representative would encounter, plus the associated clerical work. Performance measures obtained from the work sample were: record preparation, a comparison of records prepared to a model set of records; verbal contact, a sum of the ratings of verbal interaction with the customers; and filing, the sum of records not in the proper location when the work sample was ended. A composite criterion was also used, consisting of the sum of the paper-and-pencil test, record preparation, and verbal contact, minus filing, since filing was in effect an error score.

Gael and Grant (1972) found that six of the seven predictors were significantly related to the composite score for the total and the nonminority samples (number comparison was not predictive), and that three tests were significantly related for the minority sample. The Bell System Qualification Test I, number transcription, and role-play interview were significantly related to the composite score for both the minority and nonminority samples. The Bell System Qualification Test I was the best predictor of both the paper-and-pencil criterion alone ($r = .40$), and the composite score ($r = .33$).

A composite predictor score consisting of the Bell System Qualification Test I, number transcription, and the role-play interview score was compared to the composite criterion score. A multiple correlation of .37 was obtained for the total sample, with a multiple correlation of .39 for the nonminority sample and .28 for the minority sample. Regression line slopes and intercepts for the two samples were not significantly different, indicating that the composite predictor was unbiased.

The studies by Gael et al. (1975a, 1975b) used work samples to validate 10 tests of intellectual ability and perceptual speed for two occupations, telephone operators and clerks. The 10 predictors used in both studies were: The Bell System Qualification Test I, spelling, number comparison, arithmetic,

number transcription, filing, perceptual speed (circling pairs of like numbers that appear together in rows of a 40 x 25 matrix of random digits), area codes (a table of cities and area codes is presented along with a randomly arranged list of cities; the task is to associate correct area codes with the randomly listed cities), marking (numbered boxes must be marked that correspond to a 10-digit telephone number appearing above the boxes), and coding (sets of three letters are presented, and a code must be associated with each set, depending on whether the three letters are the same, whether two are the same, or all are different).

For the telephone operators (Gael et al., 1975a), the work sample consisted of handling a steady stream of incoming calls for one hour. Each activity to be performed on each call was listed on an evaluation form. Supervisors observed the subjects and underlined each activity performed incorrectly, not in accordance with trained procedures, or not at all. In addition, the overall effectiveness of each call was rated on a 5-point scale. A composite criterion was used that included the proportion of activities correct (a ratio of activities completed correctly to activities completed), cumulative work units (the number of calls completed and the complexity of the call-associated activities), and the average rating per call (averaging the 5-point ratings assigned to each call processed).

Gael et al. (1975a) compared the mean scores for white and black telephone operators on all measures. The white sample had a significantly higher mean score on every predictor but filing, and significantly higher mean scores on each criterion and the composite criterion. Every predictor was significantly related with the composite criterion for both the black and white samples. In comparing the two samples, the authors found that a common regression line overpredicts black operator proficiency and underpredicts white operator proficiency for scores below the total sample composite predictor mean. This study, then, found that the paper-and-pencil tests were valid predictors of work sample performance for both white and black operators, and that the possibility of adverse impact is more likely for nonminority than for minority candidates.

In the second study involving clerical positions (Gael et al., 1975b), the same 10 predictors mentioned above were used. Eight separate tests comprised the work sample: filing, classifying, posting, checking, coding, toll fundamentals, punched card fundamentals, and plant repair service. The first five tasks involved a variety of standardized forms that were to be processed according to specific instructions, while the last three tasks were programmed instruction booklets typically used in training courses for certain clerical jobs. The eight scores were standardized and combined into an overall proficiency score for each subject. Black, Spanish-surnamed, and white samples of newly hired clerical employees were used in the study.

In examining the mean scores for each sample on the predictors and

criteria, it was found that whites scored significantly higher than blacks on 17 of 19 measures, whites scored higher than Spanish-surnamed subjects on 12 of 19 measures, and on 7 of 19 measures Spanish-surnamed subjects obtained a higher mean than black subjects. The number of significant validity coefficients differed to a small degree across the three samples. For blacks, all 10 predictors were significantly correlated with the composite criterion, and for the white sample, only number comparison was not significantly related to the criterion. Arithmetic, number comparison, and perceptual speed were not predictive of the criterion for the Spanish-surnamed sample. Gael et al. (1975b) correlated the composite predictor of the Bell System Qualification Test I, filing, area codes, and marking with the composite criterion. The three sample regression lines had significantly different intercepts, but the slopes were not significantly different. The total sample regression line does not underpredict minority criterion scores. The authors conclude that the composite predictor is highly valid for all samples and that success in clerical work seems best predicted by tests of intellectual ability and perceptual speed and accuracy.

Performance Ratings Validated Against Work Samples

Using a slight procedural twist, Olson et al. (1981) use a functional job analysis (Fine and Wiley, 1971) to develop a work sample test for heavy equipment operators. This work sample was then used to validate the performance ratings of 360 operators, who were divided into four skill levels: high, average, low, and apprentice. The operators were tested on five different pieces of equipment, and required to perform a number of tasks on each. It was found that about 80 percent of the work sample tasks discriminated among the pre-judged operator skill levels.

Behavior Modeling Measured by Work Sample Performance

A number of studies of behavior modeling have used role playing as a form of work sample to evaluate the effectiveness of training. Moses and Ritchie (1976) had 90 managers receive supervisory relations training in such problems as reducing absenteeism, providing performance feedback, quality and quantity of work produced, insubordination, and handling discrimination complaints. A second group of 93 supervisors matched on biographical variables received no training. The work sample then consisted of three problems. Two were related to trained material: excessive absence and a discrimination complaint. The third problem, a case of suspected theft, was designed to test transfer and application of concepts learned in training. It was found that the trained group's performance on each of the three tasks was rated as significantly higher than the group that received no training. In

a similar study, Burnaska (1976) trained 62 managers in nine interpersonal skills areas. These 62 trained managers, and an additional 62 managers who had not attended the interpersonal skills training course, were then evaluated 1 month and 5 months after training. The evaluation was in the form of role play with three problems: a performance problem discussion, a work assignment discussion, and giving recognition to an average employee. A judge took the role of the employee and rated the manager on four dimensions: maintaining the employee's self-esteem, establishing open and clear communication, maintaining control of the situation, and accomplishing the objective of the discussion. Trained managers outperformed untrained managers for all three problems both at the 1- and 5-month evaluations, and the 5-month ratings were higher than the 1-month ratings.

ASSESSMENT CENTERS

An assessment center is a process that uses multiple techniques for evaluating employees for selection, promotion, placement, or special training and development (Thornton and Byham, 1982). The technique has generally been applied to managerial jobs. It seems to have its greatest value when the participant is being considered for a position very different from the one currently held, since the assessment center allows for the evaluation of skills that may not be available from observation on the current job.

Individuals are usually assessed in groups, and the assessment center staff usually consists of trained management personnel, professional psychologists, or both. The ratio of assessees to staff is usually low. The techniques used in an assessment center allow for evaluation of the assessee individually and in settings involving peer interaction.

Assessment techniques of course differ from center to center, but Thornton and Byham (1982), in reviewing approximately 500 centers, found that in-basket exercises are used by 95 percent of the centers. Some other assessment exercises and their frequency of use include: assigned-role leaderless group discussions (85 percent), interview simulations (75 percent), nonassigned-role leaderless group discussions (45 percent), management games (10 percent), reading, math, and personality tests (1 percent).

Several of the studies mentioned earlier used assessment center kinds of evaluation techniques, but for the most part there was only one type of exercise (interview simulations). A study by Petty (1974) used a leaderless group discussion as the criterion measure of the effect of training and experience on initiating structure in the group, consideration toward others in the group, and overall effectiveness in the group. One hundred ROTC students were assigned to one of four experimental conditions: experience and training (participation in a leaderless group discussion plus a 15-minute lecture on the leaderless group discussion), experience and no training, training and no experience, and

no experience and no training. Several days later the subjects were randomly assigned to groups of four students who were to discuss and prepare a complete plan of attack for an offensive tactics problem. Each group was observed by two senior ROTC students, who rated each subject on a 14-item behavioral checklist. Petty (1974) found that the training effect was statistically significant across all three criteria, but that experience had no effect. There was a significant interaction effect for training and experience on overall effectiveness. It was also found that most of the variance in the overall effectiveness rating was accounted for by the initiating structure score. Consideration provided only a negligible increase in variance accounted for, and Petty asserts that the leaderless group discussion is therefore more a measure of initiating structure than of consideration.

Ritchie and Boehm (1980) evaluated the use of biographical data and personality tests as prescreening devices for an assessment center. Eighty assessees completed a biographical information form, the Gordon Personal Profile, and the Gordon Personal Inventory. Of 10 subscales, 7 correlated significantly with final assessment center ratings. The authors applied a composite of the prescreening scores to another sample to estimate the pass rate that could be expected from using the pretests. It was found that the pass rate was raised from 44 to 48.5 percent, which would save an estimated $80,000 a year in assessment center operation costs.

MEASUREMENT ISSUES

Few work sample studies report the reliability of behavioral observations obtained by raters. In many cases, only a single rater or observer evaluates testees, and rarely is there a follow-up evaluation to provide any kind of test-retest reliability. However, three studies do address the issue.

Reliability

In Petty's (1974) study, two senior ROTC students observed each of the four subject leaderless group discussions. Each of the subjects participated in two leaderless group discussions, about 2 days apart. Three criterion measures were obtained: initiating structure, consideration, and an overall effectiveness rating. Split-half reliabilities for initiating structure and consideration were .90 and .77, respectively. Test-retest reliabilities were .62 for initiating structure, .38 for consideration, and .57 for overall effectiveness. Interrater reliabilities for the first leaderless group discussion were .74 for initiating structure, .54 for consideration, and .71 for overall effectiveness. For the second leaderless group discussion, interrater reliabilities were .65, .23, and .65 for the initiating structure, consideration, and overall effectiveness criteria, respectively.

Moses (1973) used prediction of managerial potential in a 2-day assessment center to validate a 1-day assessment center. The 2-day assessment center had previously been shown to be valid. Correlations exceeding .70 were obtained between the 2- and 1-day assessment center evaluations, indicating that the shorter, less expensive center could be used as a substitute for the 2-day center.

A laboratory study of 60 undergraduates examined the concurrent and predictive validity of a simple work sample task, as well as its test-retest reliability (Mount et al., 1977). Three predictors were used: assembling a 40-piece erector set model, the Bennett Test of Mechanical Comprehension, and the Wonderlic Personnel Test. The criterion was assembling an 80-piece erector set model. The concurrent validity group built the 40- and 80-piece models in a single session; the predictive validity group built the 40-piece model, and 9 weeks later built the 80-piece model; and the test-retest group built the 40-piece model and 9 weeks later built the 40-piece model again. For the concurrent validity group, the work sample and criterion were significantly correlated, but neither of the paper-and-pencil tests correlated significantly with the criterion. The work sample and the Bennett were both predictive of the criterion measure (.67 and .62, respectively). Finally, there was a test-retest reliability of .86 for the work sample and the criterion.

A metal trades skills work sample was designed by Schmidt et al. (1977) to emphasize oral over written instructions and tests. For the performance criteria of total tolerance and total finish, interrater reliabilities were .95 and .89, respectively. Coefficient Alpha for total tolerance was .50, for total finish was .59, and for the criterion of total work speed was .61.

Response Formats for Work Sample Evaluations

In general, work sample evaluations can use three types of response formats. Global ratings are very general evaluations of behavior and are usually on a Likert-type scale with anchors such as "performs safely or unsatisfactorily" or "performs very well or better than expected." These global ratings can be for a number of specific tasks within a work sample or for the work sample as a whole. Quite often evaluations of individual tasks are summed to obtain an overall evaluation. But again, these ratings are very nonspecific and not necessarily tied to specific behaviors observed. Assessment centers and work samples that have a pass/fail criterion quite often use this technique.

A second type of response format used in work sample evaluations is behavioral recording forms. These allow the assessor to rate work sample performance using specific examples of good and poor task behavior. Anchors are developed by job experts and indicate the specific tasks that a

testee must perform. The rater then makes a judgment as to what degree the testee exhibited the behavior required. While more specific than global ratings, this response format still requires the rater to make an evaluation along some continuum of performance. This response format is probably the most common and was used in studies by Olson et al. (1981), Frank and Wilcox (1978), and Reilly et al. (1979), for example.

A third type of response format is behavioral checklists. These are distinct from the global ratings or behavioral recording forms because the rater describes rather than evaluates the testee's behavior. A standardized checklist is developed that consists of scoring weights for each behavior, and the behaviors are particularly observable and independent of other behaviors. This method is most applicable to jobs that have a definite sequence of steps that must be performed in a particular task. This method was used by Campion (1972) in developing a work sample test for mechanics.

ADVERSE IMPACT

Work samples appear to have less adverse impact against minority groups than do paper-and-pencil tests (Howard, 1983). Two studies directly compared the adverse impact of work sample predictors to paper-and-pencil predictors. Field et al. (1977) compared a minority sample of 52 production workers with 48 nonminority workers in a boxboard container plant. The paper-and-pencil tests used were the Personnel Tests for Industry-Numerical (Form A) and Personnel Tests for Industry-Oral Directions Test (Form S), which measured basic math skills and general mental ability, respectively. Two short work samples were designed to test use of a ruler in measuring various dimensions of a three-dimensional figure and the ability to read and decipher computer printout specifications for making a box. Two criteria were used: a supervisor performance rating, requiring the supervisor to rate each employee on six dimensions of the job; and a productivity measure of the number of boxboard containers produced. Field et al. found that the mean score on the four predictors and the two criteria was higher for nonminority than for minority employees. However, validity coefficients for the two samples showed no adverse impact. Both work samples were significantly related to the two criterion measures for the two samples of employees. For the two paper-and-pencil predictors, the numerical test was significantly related to the performance appraisal criterion. All other validities failed to reach statistical significance. The work samples in this study, therefore, showed no adverse impact and better predictive validity than the two paper-and-pencil tests.

Kesselman and Lopez (1979) compared a paper-and-pencil predictor (Personnel Classification Test, yielding a verbal, numerical, and total score) with a written, accounting job knowledge test. (The personnel classification test was

chosen prior to a detailed job analysis for the accountant position, while the job knowledge test was designed specifically from the job analysis. The following results must be considered with this in mind.) The two criteria were composite job knowledge proficiency, measured by an 18-item behavioral observation scale completed by the employee's supervisor, and an overall job performance rating, also completed by the employee's supervisor.

The sample of 52 accountants was analyzed according to sex (27 male and 25 female), and race (28 minority and 24 nonminority). On the job knowledge test, no significant differences were found between the means of the two ethnic groups and the male-female groups. The personnel classification test means showed differences between minorities and nonminorities and between sexes: a higher average score was obtained by whites and by males. No differences were found in the group means for the proficiency criterion, but on overall job performance, females were rated significantly higher than males.

The job knowledge test was found to be a valid predictor of the proficiency criterion for all groups except the minority sample. The job knowledge test was predictive of overall job performance only for the female sample. No validity coefficients for the personnel classification test reached significance. The job knowledge test showed no significant minority-nonminority differences for the slopes and intercepts of the regression lines for the two criteria. While the slopes for the personnel classification test are not significantly different for the two racial groups, the intercepts are different, and a common regression line would underpredict minority criterion values. Kesselman and Lopez (1979) conclude that while the paper-and-pencil predictor showed adverse impact for the minority sample, a job knowledge test carefully constructed from a job analysis eliminated this problem. It should be emphasized again, however, that Kesselman and Lopez chose the personnel classification test prior to a job analysis. The findings of this study would be more impressive had some effort been made to use a standardized predictor that tapped abilities and aptitudes uncovered by the job analysis.

Grant and Bray (1970) found that slopes of the regression lines were equal for minority and nonminority telephone company repairmen in a job training situation, and that the difference in intercepts was actually slightly biased against nonminority candidates. Arnold et al. (1982) found that a strength test for selecting steelworkers would have, at most, a slight adverse impact against males.

The series of studies by Gael, Grant, and Ritchie (Gael and Grant, 1972; Gael et al., 1975a, 1975b) all specifically compared the validities of paper-and-pencil predictors and work sample criteria for minority and nonminority employees. All three studies found that nonminority employees scored significantly higher than minority employees on the predictor and the criterion measures, but that validity coefficients were comparable. Also, in each case

the authors found that a common regression line did not underpredict minority employee proficiency.

In a study of 87 metal trades apprentices, Schmidt et al. (1977) found that all five subscores of a written job knowledge test (Machine Trades Achievement Test) showed large differences between minorities and nonminorities. However, two of three job sample subscores (tolerance and finish) that required completing a workpiece with oral instructions showed no significant subgroup differences.

Cascio and Phillips (1979) compared white, black, and Latin raters and ratees on 10 verbal and 11 motor work sample tests. By systematically training raters, clearly defining performance standards, and using content valid tests, the authors found average interrater reliabilities of .93 for promotional tests, .87 for entry level tests, .91 for motor tests, and .89 for verbal tests. No evidence for disparate impact was found for any of the rater and ratee race combinations, leading Cascio and Phillips (1979) to term performance tests as "a rose among thorns."

Brugnoli et al. (1979) examined racial bias in a work sample test for maintenance mechanics. Fifty-six white, male maintenance mechanics evaluated a videotape of a black job applicant and a white job applicant performing a relevant task of laying out, drilling, and tapping, and an irrelevant task of indexing drill bits. The raters then used a highly specific behavioral recording form, a global evaluating form, or both. The only condition in which bias was found involved global evaluations of irrelevant behavior. No bias was found when a behavioral recording form or a global form was used for relevant behaviors, nor for global evaluations made following behavioral recordings. The authors conclude that work samples based on performance that is critical to success or failure on the job, especially when combined with behavioral recordings, will have little potential for racial bias.

One study that did find race and sex bias in a work sample was reported by Hamner et al. (1974). Undergraduate college students acting as managers rated all eight combinations of male/female and black/white job performers. This laboratory study's work sample task was stocking a grocery shelf with large cans. Performance was systematically varied: high performers stocked 48 cans in 3 minutes, while low performers stocked 24 cans in 3 minutes. Global performance was rated on a 15-point scale ranging from weak in overall performance to exceptionally good in overall performance. It was found that 30 percent of the variance in ratings was due to performance, but the higher ratings were given to performers of the same race, higher ratings were given to females, high performing females were rated higher than high performing males, high performing blacks were rated only slightly higher than low performing blacks, while high performing whites were rated much higher than low performing whites. Twenty-three percent of the variance in ratings was due to sex/race combinations. While this study did find some

instances of race and sex bias, it was an extremely simple task using global evaluation. Both of these study characteristics could explain the results, especially in light of Brugnoli et al.'s (1979) study.

It appears that work sample tests offer an opportunity for reducing adverse impact while at the same time obtaining comparable or even better validities than more traditional predictors or criteria. A thorough job analysis, which normally precedes the development of any predictor or criterion measure, is particularly advantageous in the case of a work sample. By tying the work sample closely to the knowledge, skills, and abilities actually required in a job, any racial differences that do appear should be no greater than actual job performance differences. This approach of course supposes that unfair bias will not enter into the performance appraisal process through global evaluations of irrelevant job behavior (Brugnoli et al., 1979).

CONCLUSION

The research concerning work sample testing suggests that they can produce high predictive validities, and that when used as criteria they compare favorably with supervisor ratings and productivity measures. Work samples appear to be particularly relevant in training situations, as both a measure of training success and as a means of assessing the trainability of individuals prior to a full-length training program. Also of considerable importance is the fact that work sample tests seem to reduce adverse impact, particularly if the ratings concentrate on relevant job tasks. In the few studies that address reliability issues, work samples show good test-retest and interrater reliabilities.

Unfortunately, a large gap exists in the literature with regard to work samples as measures of incumbent employees' performance. This use of work samples as a criterion measure apart from other forms of performance appraisal may be beneficial, however. In jobs that require a high degree of specific technical skills, or in which a core of critical job behaviors can be identified, work samples would be an additional method of obtaining performance data. If properly constructed, they eliminate rating biases by requiring the evaluator to describe the employee's behavior on a standardized form rather than to evaluate the behavior observed. This may reduce or eliminate some of the more common rating errors, such as halo, leniency, or central tendency, since the rater's only judgment is whether a behavior has in fact occurred, and not to what degree or how appropriate that behavior is.

Because of their standardized nature, and the fact that rating occurs while the behavior is taking place, work samples are less prone to the errors arising from a time lag between observation and rating. As mentioned early in this paper, because of their close tie to actual work behaviors, work

samples also allow an interaction of abilities and skills to occur, an interaction that is often artificially eliminated by rating forms with generalized dimensions of work behavior.

Work samples, however, are not directly substitutable for all forms of performance appraisal. While appropriate for testing certain skills, other, more traditional supervisory evaluations may provide data about interpersonal skills, initiative, etc. if they are indeed critical for job success. These types of skills may be evaluated in an assessment center setting. However, the direct link between assessment center behaviors and job behaviors might not be as clear as the link between motor work samples, for example, and job behaviors. Nonetheless, work sample evaluations can provide an additional source of criterion data that can be thought of as more objective and standardized than supervisory performance ratings. In fact, as Dunnette and Borman (1979) note, we should perhaps not expect high agreement between differing sources of performance evaluation information. While traditional organizational structure makes it the responsibility and even the right of the supervisor to evaluate his or her employees, this practice does not automatically define the supervisor's view as reality. Multiple sources of criterion data should more accurately define an employee's performance, and work samples appear to be extremely useful in this regard.

REFERENCES

Arnold, J.D, J.M. Rauschenberger, W.G. Soubel, and R.M. Guion 1982 Validation and utility of a strength test for selecting steelworkers. *Journal of Applied Psychology* 67:588-604.

Asher, J.J. 1972 The biographical item: can it be improved? *Personnel Psychology* 25:251-269.

Asher, J.J., and J.A. Sciarrino 1974 Realistic work sample tests: a review. *Personnel Psychology* 27:519-533.

Brugnoli, G.A., J.E. Campion, and J.A. Basen 1979 Racial bias in the use of work samples for personnel selection. *Journal of Applied Psychology* 64:119-123.

Burnaska, R.F. 1976 The effects of behavior modeling training upon managers' behaviors and employees' perceptions. *Personnel Psychology* 29:329-335.

Campion, J.E. 1972 Work sampling for personnel selection. *Journal of Applied Psychology* 56:40-44.

Cascio, W.F., and N.F. Phillips 1979 Performance testing: a rose among thorns? *Personnel Psychology* 32:751-766.

Cohen, S.L., and L.A. Penner 1976 The rigors of predictive validation: some comments on "A job learning approach to performance prediction." *Personnel Psychology* 29:595-600.

Dunnette, M.D., and W.C. Borman 1979 Personnel classification systems. *Annual Review of Psychology* 30:477-525.

Field, H.S., G.A. Bayley, and S.M. Bayley 1977 Employment test validation for minority and nonminority production workers. *Personnel Psychology* 30:37-46.

Fine, S.A., and W.W. Wiley 1971 *An Introduction to Functional Job Analysis*. *Methods for Manpower Analysis*. Monograph No. 4. Kalamazoo, Mich.: W.E. Upjohn Institute.

Fleishman, E.A. 1964 *The Structure and Measurement of Physical Fitness*. Englewood Cliffs, N.J.: Prentice-Hall.

Frank, H., and C. Wilcox 1978 Development and preliminary cross-validation of a two step procedure for firefighter selection. *Psychological Reports* 43:27-36.

Gael, S., and D L. Grant 1972 Employment test validation for minority and nonminority telephone company service representatives. Journal of Applied Psychology 56:135-139.

Gael, S., D.L. Grant, and R.J. Ritchie 1975a Employment test validation for minority and nonminority telephone operators. Journal of Applied Psychology 60:411-419.

1975b Employment test validation for minority and nonminority clerks with work sample criteria. Journal of Applied Psychology 60:420-426.

Gordon, M.E., and L.S. Kleinman 1976 The prediction of trainability using a work sample test and an aptitude test: a direct comparison. *Personnel Psychology* 29:243-253.

Grant, D.L., and D.W. Bray 1970 Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology* 54:7-14.

Hamner, W.C., J.S. Kim, L. Baird, and W.J. Bigoness 1974 Race and sex as determinants of ratings by potential employers in a simulated work sampling task. *Journal of Applied Psychology* 59:705-711.

Howard, A. 1983 Work samples and simulations in competency evaluations. *Professional Psychology: Research and Practice* 14:780-796.

Hunter, J.E., and R.F. Hunter 1984 Validity and utility of alternative predictors of job performance. *Psychological Bulletin* 96:72-98.

Hunter, J.E., F.L. Schmidt, and R. Hunter 1979 Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin* 86:721-735.

Inskeep, G.C. 1971 The use of psychomotor tests to select sewing machine operators: some negative findings. *Personnel Psychology* 24:707-714.

Kesselman, G.A., and F.E. Lopez 1979 The impact of job analysis on employment test validation for minority and nonminority accounting personnel. *Personnel Psychology* 32:91-108.

Moses, J.L. 1973 Assessment center for the early identification of supervisory and technical potential. In W.C. Byham and D. Bobin, eds., *Alternatives to Paper and Pencil Testing*. Proceedings of a conference at the Graduate School of Business, University of Pittsburgh.

Moses, J.L., and R.J. Ritchie 1976 Supervisory relations training: a behavioral evaluation of a behavior modeling program. *Personnel Psychology* 29:337-343.

Mount, M.K., P.M. Muchinsky, and L.M. Hanser 1977 The predictive validity of a work sample: a laboratory study. *Personnel Psychology* 30:637-645.

Olson, H.C., S.A. Fine, D.D. Myers, and M.C. Jennings 1981 The use of functional job analysis in establishing performance standards for heavy equipment operators. *Personnel Psychology* 34:351-364.

Petty, M.M. 1974 A multivariate analysis of the effects of experience and training upon performance in a leaderless group discussion. *Personnel Psychology* 27:271-282.

Reilly, R.R., and W. Manese 1979 The validaton of a minicourse for telephone company switching technicians. *Personnel Psychology* 32:83-90.

Reilly, R.R., S. Zedeck, and M.L. Tenopyr 1979 Validity and fairness of physical ability tests for predicting performance in craft jobs. *Journal of Applied Psychology* 64:262-274.

Ritchie, R., and V. Boehm 1980 Reducing costs by prescreening assessment center candidates. *Assessment and Development* 7(2):5.

Robertson. I., and S. Downs 1979 Learning and the prediction of performance: development of trainability testing in the United Kingdom. *Journal of Applied Psychology* 64:42-50.

Robertson, I.T., and R.S. Kandola 1982 Work sample tests: validity, adverse impact, and applicant reaction. *Journal of Occupational Psychology* 55:171-183.

Robertson, I.T., and R.M. Mindel 1980 A study of trainability testing. *Journal of Occupational Psychology* 53:131-138.

Schmidt, F.L., A.L. Greenthal. J.E. Hunter, J.G. Berner, and F.W. Seaton 1977 Job sample vs. paper-and-pencil trades and technical tests: adverse impact and examinee attitudes. *Personnel Psychology* 30:187-197.

Schmitt, N., R.Z. Gooding, R.A. Noe, and M. Kirsch 1984 Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology* 37:407-422.

Siegel, A.I. 1983 The miniature job-training and evaluation approach: additional findings. *Personnel Psychology* 36:41-56.

Siegel, A.I., and B.A. Bergman 1975 A job learning approach to performance prediction. *Personnel Psychology* 28:325-339.

Thornton, G.C., III, and W.C. Byham 1982 *Assessment Centers and Managerial Performance*. New York: Academic Press.

Wernimont, P.F., and J.P. Campbell 1968 Signs, samples, and criteria. *Journal of Applied Psychology* 52:372-376.

Measuring Job Competency

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

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 within the National Academy of Science's National Research Council 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

This paper was produced to reflect the joint discussions of the Committee on the Performance of Military Personnel and the Job Performance Measurement Working Group that took place on October 24-25, 1986, and March 13-14, 1987.

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 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

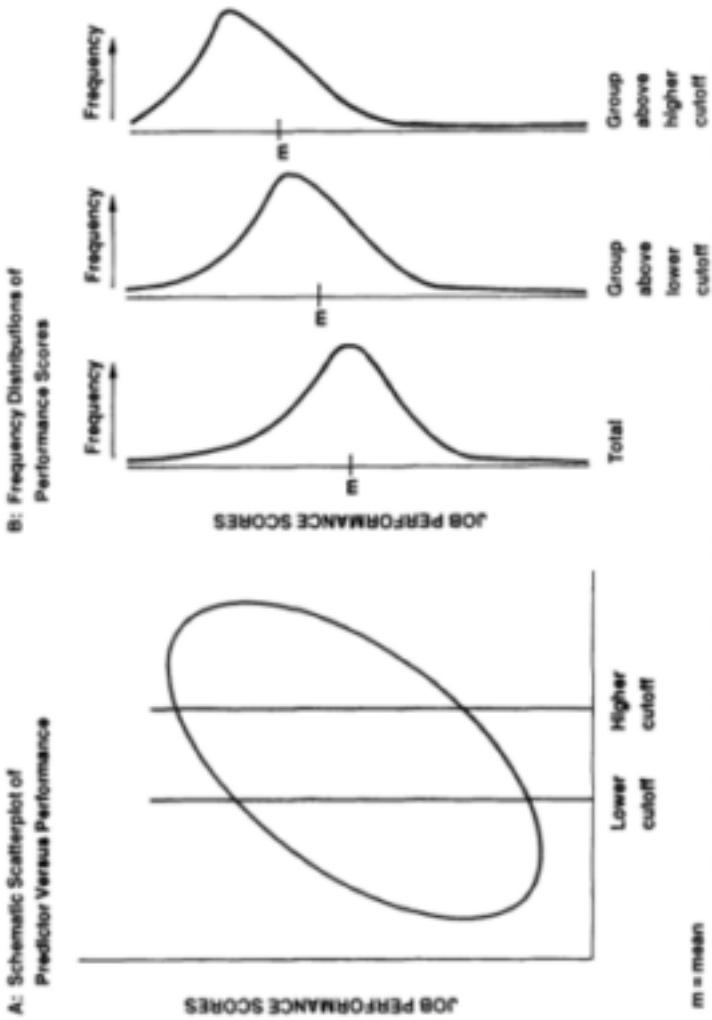


Figure 1 Scores on predictor composite with resultant performance distributions.

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

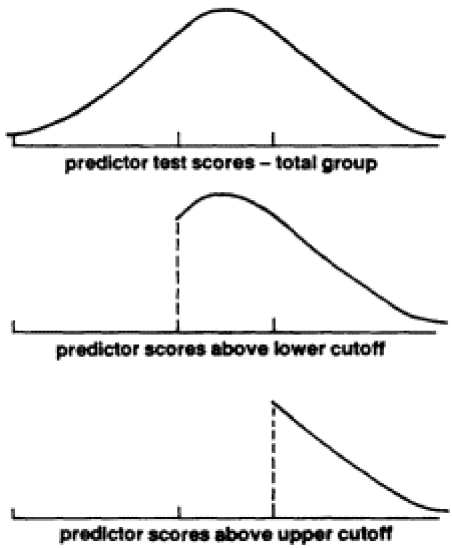


Figure 2 Distribution of predictor composite scores.

distribution is at issue, not just some acceptable minimum performance level.[1]

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 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

[1] 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.

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 be 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 important. 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 competency 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 paper 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.[2] The zero point could be defined as absence of performance, i.e., being present but with no activity.

[2] 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

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.

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.

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).

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.

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.

The Evaluation of Alternative Measures of Job Performance

Linda S. Gottfredson

INTRODUCTION

The Criterion Problem in Personnel Research

The "criterion problem" is one of the most important but most difficult problems in personnel research. One book on the theory and methods of performance appraisal (Landy and Farr, 1983:3) referred to the measurement of job performance as still one of the most vexing problems facing industrial-organizational psychologists today despite over 60 years of concern with the topic. It is a vexing problem because job performance can be measured in many ways, and it is difficult to know which are the most appropriate, because there is generally no empirical standard or "ultimate" criterion against which to validate criterion measures as there is for predictor measures. One need only ask a group of workers in the same job to suggest specific criterion measures for that job in order to appreciate how difficult it is to reach consensus about what constitutes good performance and how it can be measured fairly.

The criterion problem is important because the value of all personnel policies from hiring to promotion and employee counseling depends on the appropriateness of the job performance standards to which those policies

I gratefully acknowledge the critical comments made on earlier drafts of this paper by Bert F. Green, Jr., Robert M. Guion, Frank J. Landy, Frank L. Schmidt, and Alexandra K. Wigdor.

are tied. For example, no matter how well one's selection battery predicts later criterion performance, that battery may do little good for the organization if the job performance criterion measure against which it was validated is inappropriate. Personnel researchers have often been criticized for seizing the most available criterion measure (Jenkins, 1946; Guion, 1961), and as a result, more research has been devoted in recent decades to developing new and more elaborate types of performance measures (for example, behaviorally anchored rating scales and work samples). However, our understanding of the relative strengths and weaknesses of different classes of criterion measure is still meager enough that Wherry's (1957:1) comment three decades ago still is all too apt: "We don't know what we are doing, but we are doing it very carefully . . . ."

The literature on the criterion problem has provided some general standards by which to classify or evaluate job performance criterion measures, such as closeness to organizational goals, specificity, relevance, and practicality (e.g., Smith, 1976; Muckler, 1982). But the literature also reflects a history of debate about the proper nature and validation of a criterion measure (e.g., Wallace, 1965; Schmidt and Kaplan, 1971; James, 1973; Smith, 1976). For example, should criterion measures be unidimensional? If somewhat independent dimensions of job performance are measured, perhaps multiple rather than composite criteria are indicated. Should the aim be to measure economic or behavioral constructs, and what role do construct and content validation methods play in validating such measures? Is it necessary for the criterion measure to mimic tasks actually performed on the job? Should measures be general or specific in content? And when must they be criterion-referenced rather than norm-referenced? Different classes of measures, such as global ratings, behaviorally anchored rating scales, work sample tests, and paper-and-pencil job knowledge tests have been discussed at length.

What these debates illustrate is that there are many possible criterion measures, that all measures have drawbacks, and that it is largely the organization's goals for criterion measurement that determine which measures are most appropriate in given situations. The question "criteria for what?" therefore has been a useful guide to criterion evaluation, but a researcher seeking more specific guidelines from the literature for validating (rather than constructing) a criterion measure will be disappointed.

Besides serving as criteria for validating personnel selection and classification procedures, job performance measures can serve diverse other purposes: for example, feedback to individuals, redirecting worker behavior, human resource planning, and decisions on how to carry out training, promotion, and compensation. The term "performance appraisal" is usually used to designate these latter administrative purposes. The same measures often have different advantages and disadvantages, depending on the organization's

particular goal for measuring job performance, but issues in the evaluation of job performance measures are basically the same whether those measures are used for validating predictors or for the other purposes just listed. Thus, although this paper focuses on evaluating job performance measures in their role as criteria in developing personnel selection procedures, it has more general applicability.

In this paper some strategies are suggested for evaluating criterion measures. It will be evident to the reader, however, that the criterion problem is a web of problems ready to ensnare even the most able and dedicated explorers of the criterion domain.

Evolution of the Criterion Problem

The dimensions of the criterion problem in its current manifestations can be appreciated by reviewing the evolution of criterion problems in personnel research. The field of personnel research was born early in this century as employers tried to deal with severe job performance problems such as high accident rates in some industries and phenomenal turnover rates by today's standards in many others (Hale, 1983). Criterion measures leapt out at employers, and the need in personnel research was to find predictors of those worker behaviors and to help employers develop coherent personnel policies.

A plethora of employment test batteries was subsequently developed for use in industry. Both military and civilian federal agencies provide examples of systematic research programs begun early in this century to develop and validate test batteries for the selection and classification of employees. The General Aptitude Test Battery (GATB) (U.S. Department of Labor, 1970) is a product of the U.S. Employment Service and the Armed Services Vocational Aptitude Battery (ASVAB) (U.S. Department of Defense, 1984) is the latest generation test battery developed by the military for selection and classification.

By mid-century the search for predictors had led not only to the development of a variety of useful personnel selection devices, but it had also produced hundreds of predictive validity studies. The accumulation of these studies began to make clear that much greater care was being given to the development of predictors than to the criterion measures against which they were being validated. Discussions of the criterion problem began to appear with increasing frequency (e.g., Jenkins, 1946; Brogden and Taylor, 1950: Severin, 1952; Nagle, 1953; Wherry, 1957; Guion, 1961; Astin, 1964; Wallace, 1965) and the profession turned a critical eye to the problem. The result of that concern has been a search for criterion measures that may some day rival the earlier and continuing search for predictors.

Commonly used criterion measures received considerable criticism. Performance

in training had been (and still is) commonly used to validate predictor batteries, as is illustrated by the manuals for both the GATB and the ASVAB (U.S. Department of Labor, 1970; U.S. Department of Defense, 1984). But training criteria were increasingly criticized as being inappropriate substitutes for actual job performance where the aim was, in fact, to predict future job performance (e.g., Cronbach, 1971:487). This was particularly the case after Ghiselli (1966) compiled data showing differential predictability for training versus on-the-job performance measures. The ubiquitous supervisor rating was considered too subject to rater subjectivity; on the other hand, most objective measures such as production records or sales volume were criticized as being only partial measures of overall performance and as being contaminated by differences in working conditions not under the worker's control.

These criticisms have been accompanied by efforts to improve existing measures as well as to develop new ones. Ratings have been the object of considerable research, and several theoretical models of the rating process (Landy and Farr, 1983; Wherry and Bartlett, 1982) have been produced to guide the design of better rating scales. Evidence suggesting that job performance is complex and multidimensional led to discussions of when multiple criteria are more useful than composite criteria and of how the components of a composite criterion should be weighted (Nagle, 1953; Guion, 1961; Schmidt and Kaplan, 1971; Smith, 1976). New types of rating scales—in particular, behaviorally anchored rating scales—were designed with the intention of overcoming some of the inadequacies of existing rating scales, and work sample tests have attracted considerable attention in recent years with their promise of providing broad measures of performance with highly relevant test content.

The search for better measures of job performance has not been entirely the outgrowth of professional research and debate, but has been driven in no small part by social, economic, and political forces. For example, sociolegal standards for assuring fairness in personnel policies have become more demanding in recent years and require that organizations adopt the most highly job-related selection tests if their selection tests have adverse impact on some protected group. This in turn has stimulated a greater demand for valid performance criterion measures to establish job-relatedness.

Although the military is not subject to the same equal employment opportunity regulations as are civilian employers, its current personnel research activities illustrate yet other pressures for the development of new or better measures of job performance: specifically, the need to assess and increase the utility of personnel policies (e.g., see Landy and Farr, 1983:Ch. 9). For example, personnel selection and classification procedures have become of increasing concern because the eligible age cohort for military recruitment will be shrinking in size in the coming years, which means that

the military has to make the best possible use of the available pool of applicants. In addition, the quality of the applicant pool has fluctuated to reach uncomfortably low levels in recent years (e.g., see Armor et al., 1982: Figure 1) and may do so again in the future, while at the same time military jobs are becoming increasingly complex. A frequently expressed concern in this regard is that the military, like many civilian employers, may be wasting nonacademic talent by validating predictors against academic criteria such as training grades when jobs themselves may not depend so heavily on verbal ability or academic skills. It must be recognized that trainability is itself important because of the high costs associated with training. Nevertheless, validating predictors against direct measures of job performance might reveal that there are more qualified applicants for some military jobs than has appeared to be the case in the past. If this were the case, mission effectiveness might be sustained or even improved despite a more limited recruit pool if that pool were utilized more efficiently.

In short, past job performance measures have been useful, but there has been constant pressure from within and from outside the research community to improve and expand the measurement of job performance and thereby improve the utility of all personnel policies based on such measures. Related developments, such as improved computer technology for handling large data bases and the development during the last two decades of task analysis methods and data, which are required for building certain job performance measures, have also improved prospects for developing sound measures of job performance.

The current state of the criterion problem is illustrated by the efforts of the U.S. military's Job Performance Measurement Project (JPM) for linking enlistment standards to on-the-job performance (Office of the Assistant Secretary of Defense, 1983). In its effort to develop good job performance criteria for validating enlistment standards, that project is developing and evaluating at least 16 distinct types of job performance criterion measures: 7 measures of performance on specific work tasks (e.g., work samples, computer simulations, task ratings by supervisors) and 3 sources each for performance ratings on task clusters, behavior dimensions, and global effectiveness. These measures differ considerably in specificity and type of item content, who evaluates performance, and the stimulus conditions presented to examinees.

Although no claim is made that these JPM measures will all measure exactly the same thing, they are being investigated as possible alternative measures of the same general performance construct (technical proficiency) for exactly the same use (validating selection and classification procedures in the four Services). Ostensibly, the evaluation issue is not one of choosing one kind of job performance construct over another or of finding some optimal composite of different dimensions of performance, as has been the case in past discussions of specific and quite different performance criteria

such as quantity of work, number of errors, absenteeism, salary, or promotion rate. Research and development have proceeded to the point where we now have a variety of viable contenders for the title of "best overall measure of job performance of type X for purpose Y."

The JPM Project vividly illustrates that the search for new and better criterion measures has led the field to a new frontier in the criterion problem, one that arises from the luxury of choice. Namely, how should alternative measures that were designed to serve the same purpose be evaluated and compared, and by what standards should one be judged more useful or appropriate than another for that purpose?

The objective of this paper is to outline the major issues involved in evaluating alternative measures of the same general type of performance to be used for the same purpose. At the outset, however, it is important to note that this task actually differs only by degree from the task of evaluating and selecting from among measures of distinctly different kinds of performance. Realistically, even measures that have been designed to measure exactly the same thing are unlikely to do so; instead, they can be expected to measure at least somewhat different facets of performance—some desired and some not. Moreover, general measures of technical proficiency, such as work samples and supervisor ratings, are usually presumed to measure different specific, but unspecified, types of proficiency and to different degrees (Vineberg and Joyner, 1983). Thus, as will be discussed in detail later, selecting among different measures of the same general type of performance is likely, in fact, to involve making a choice among meaningfully different kinds of performance.

This new aspect of the criterion problem is often referred to as the investigation of *criterion equivalence*. I will adhere to this common terminology, but it should be clear that equivalence versus nonequivalence is not the issue. The issue is one of type and degree of similarity.

THE NATURE OF CRITERION EQUIVALENCE

Measures of job performance—even obvious criteria—should be systematically evaluated before an organization adopts any of them. If the organization fails to evaluate its potential alternative measures explicitly and carefully, it risks adopting measures that do not meet its needs as well as might other alternatives.

Validity, reliability, and practicality or acceptability are the three general standards that have most often been suggested for evaluating the quality of a criterion measure (e.g., Smith, 1976; Landy and Farr, 1983). The purpose of applying such standards may be to facilitate decisions about which, if any, criterion measure will be adopted in a given setting; it may be to help improve the criterion measures under consideration; or it may

be to verify that the criterion measures that have been developed do in fact function as intended. As will be illustrated, the selection of a criterion measure (or set of measures) is ultimately a judgment about how highly the organization values different types of performance, so an explicit evaluation of alternative criterion measures can also be useful if it stimulates greater clarification of the organization's goals for the measurement of job performance.

Five Major Facets of Equivalence

Five general facets of equivalence among criterion measures are discussed below: validity, reliability, susceptibility to compromise (i.e., changes in validity or reliability with extensive use), financial cost, and acceptability to interested parties. The first two have been the issues of greatest concern to researchers. The third issue has been only implicit in previous discussions of criterion quality, but is important. The last two facets of equivalence are both types of acceptability or practicality, but they are distinguished here because they often require different responses from the organization.

Although all dimensions should be of concern to the researcher as well as to the decision makers in the organization, the organization must rely most heavily on the researcher for information about the first three. In turn, researchers must be fully apprised of the organization's goals for performance measurement, because all facets of equivalence depend on what uses will be made of the criterion measures. The evaluation of criterion measures cannot be divorced from the context of their use.

Validity

The first requirement of a criterion measure is that it actually function as intended. If the criterion measure does not measure the performances that promote the organization's goals, if it is not clear whether the measure does so or not, or if the organization's goals for measurement are unclear, then other facets of nonequivalence such as cost and acceptability are irrelevant.

Determining validity is the essence of the criterion problem, and so too is it the troublesome central issue in the comparison of any two or more measures. Moreover, what constitutes validity is a subject of considerable debate. For these reasons, the nature of validity and how it can be established is explored in detail in later sections of this paper. Briefly stated, however, validation is a process of hypothesis testing. Two types of hypotheses are of concern in the evaluation of job performance measures: (1) construct validity, which refers to inferences about what performance construct has actually been operationalized by a measure and (2) relevance,

which refers to the relation of the performance construct to the organization's goals for performance measurement, such as increased organizational effectiveness.

Reliability

From the standpoint of classical test theory, reliability is the proportion of variance in observed scores that is due to true score differences among individuals rather than to error of measurement. Estimating reliabilities can be a difficult problem, especially for criterion measures that require ratings of some sort. Generalizability theory (Cronbach et al., 1972) provides one systematic way of estimating the amount of variation associated with different sources of variation (e.g., raters, instability over time, item or subtest), one or more types of which the investigator may choose to regard as error, depending on the criteria being compared and the context of their projected use.

Although good reliability estimates are essential for making good decisions about which criterion measures to adopt, the reasons for their importance vary according to the projected uses of those measures. When workers' scores on a job performance measure are used directly in making decisions about the promotion or compensation of those workers or in providing feedback to them, then unreliability reduces the utility of the performance measure. Specifically, using a less reliable measure rather than a more reliable one (assuming that they measure the same thing otherwise) means that the organization is promoting, rewarding, or counseling workers relatively more often than need be on the basis of error in measurement rather than on the performances it values; thus, the organization is not reinforcing the desired worker behaviors as efficiently as it might. An unreliable measure of true performance levels may also be a source of much discontent among workers and supervisors (as also might, of course, a reliable but irrelevant or biased measure), which would further decrease the utility of the measure to the organization.

If a performance measure is used only as a criterion for selecting a predictor battery, unreliability does not directly affect the utility of the predictor battery selected and so neither does it affect the utility of the criterion measure itself. Assuming adequate sample sizes, a less reliable criterion measure will select the same predictor battery as will a more reliable one if the two do in fact measure the same type of performance. The only difference will be that the weights for the predictors will be proportionately lower for the less reliable criterion measure. This difference in weights is of no practical consequence because the two resulting prediction equations will select the same individuals from a pool of applicants.

However, it is not possible to determine the utility of a criterion measure to the organization or the utility of the battery for predicting criterion performances unless criterion reliability has been estimated. As discussed later, assessing the utility of a criterion measure requires a knowledge of its validity; assessing its validity requires estimates of its true score correlations with other variables; and these in turn require an assessment of reliabilities. Similarly, assessing the utility of a battery for *predicting* criterion performances requires an estimate of the correlation between observed scores on the predictor and true scores on the criterion measure, and this requires a reliability estimate for criterion scores.

Susceptibility to Compromise

Susceptibility to compromise refers to the ease with which the initial reliability or validity of the criterion measure can be damaged during extended use. Stated conversely, susceptibility to compromise refers to the difficulties or requirements the organization faces in maintaining the initial psychometric integrity of the criterion measure. What is at issue here is not the level of a criterion measure's reliability or validity, but the degree to which its initial reliability or validity is likely to *fluctuate* to some unknown degree, resulting also in changes in the proper interpretation of test scores and in the utility of the measure.

In general, the more carefully specified and constrained the examiner's behavior, the less need there is to carefully select, train, and monitor examiners. Job performance measures differ in the amount of judgment and discretion they require of examiners and so differ also in the amount of control they require over examiners if their initial psychometric integrity is to be maintained in the field over time. For example, all types of rating scales and work sample tests require examiners or raters to rate the quality of performances they observe, which leaves room for changes in levels of rater carelessness, rating halo, rater leniency and central tendency, and rater prejudices against certain types of workers—all of which are errors that decrease the reliability or the validity of criterion scores. Such criterion measures are very different from multiple-choice, paper-and-pencil job knowledge tests, because a cadre of test examiners or raters who are well trained in how to rate accurately different performance levels is required for the former but not the latter. More objectively scored tests are not necessarily immune to degradations in quality because test administration may decay in quality. For example, the enforcement of time limits may become lax or the type and number of prompts or cues given to examinees may change over time.

Test security and reactivity reflect compromises of validity stemming from examinee behavior on the test and so are concerns with all types of

criterion measures. The former refers to the bias introduced when examinees know in advance what the test items are, and it is particularly a concern with written job knowledge tests, work sample tests, and other tests of maximal performance. Breaches of test security and their consequences for job knowledge tests can be minimized by frequent test revisions, by using alternative forms, or perhaps by employing the developing technology of adaptive testing (Curran, 1983). Good logistics at the testing sites for paper-and-pencil or work sample tests can also minimize accidental as well as intentional cheating. The security problems posed by such tests can differ dramatically, however. For example, paper-and-pencil job knowledge tests can be administered en masse to examinees in a relatively short period of time, whereas work sample tests are often administered individually and the number of people tested at one time depends on the amount of equipment and the number of personnel that can be devoted to testing. This in turn means that there is much more opportunity for intentional or accidental breaches of test security of the latter than the former because individuals yet to be tested cannot be segregated for more than very short spans of time from individuals who have already been tested. Test administrators and examinees can be admonished to refrain from discussing test content with potential examinees, but it seems unrealistic to expect voluntary restraint to be effective for the days, weeks, or even months that may be required for work sample testing at some sites.

Reactivity refers to changes in performance that are simply a function of examinees knowing that they are being observed and evaluated. Reactivity influences the initial reliability and validity of a criterion measure, as does any other source of error or bias, but it also illustrates well one type of compromise of psychometric integrity. That compromise is possible when perceptions of the consequences of performance measurement change over time. For example, supervisor ratings might be developed and evaluated for research purposes but then later be adopted by the organization for evaluating employees for retention, promotion, or salary administration. Supervisors and their employees may be unconcerned about how favorably workers are evaluated when criterion measures are used for research purposes. However, they have a greater stake in the outcomes of measurement when those scores are used to punish or reward workers (and indirectly their supervisors too), and both supervisors and their employees may engage in what Curran (1983:255) has referred to as "gaming." Thus, if the supervisor ratings were originally perceived as nonthreatening by employees, but those perceptions change for some reason, then the reliability and validity of the ratings as documented in the original research probably will differ from that for subsequent use of the criterion measure. Consistent with this, Bartlett (1983) found that scores obtained twice on the same performance measure,

administered once for research purposes and then again for performance appraisal, are sometimes uncorrelated.

In short, susceptibility to compromise is not entirely an inherent feature of a criterion measure, but also depends on the uses to which the job performance measure will be put and on the steps the organization takes to maintain the initial psychometric properties of the criterion measure over time. The greatest risk of compromise accompanies the use of measures for performance appraisal, but some risk also accompanies the extended research use of a measure.

Financial Cost

The cost of developing and administering a criterion measure depends to a large extent on how carefully it is developed, how fully it is evaluated, and how well it is administered. Carefully developing and evaluating criterion measures may be a costly process regardless of type of criterion measure, and the major differences in cost may be in their administration. Work sample tests are often described as being relatively expensive in terms of equipment costs at the test sites, lost work time of examinees and their supervisors, costs of employing the additional testing personnel, and disruption to organizational operations (Vineberg and Joyner, 1983). Paper-and-pencil tests appear to be much less costly in all these respects, except perhaps when few people are to be tested (Cascio and Phillips, 1979). Ratings are relatively inexpensive to administer if they are gathered infrequently, but requiring raters to make periodic ratings on the same individual or to make notes concerning individuals that would later be used in making ratings (e.g., in an attempt to reduce illusory halo) can be costly in terms of lost supervisor time and goodwill. The costs of administering tests weigh more heavily when those measures are used for performance appraisal as well as (or rather than) occasional research purposes, because then the ratio of administration to development costs is greater.

Acceptability to Interested Parties

The direct financial costs of a criterion measure influence how acceptable it is to the organization, but it is important to identify other types of acceptability that may have only indirect financial consequences. These include the acceptability or legitimacy of the criterion measure to other interested parties, including the workers being evaluated, their unions, supervisors who may be responsible for collecting data, professional organizations, and funding or regulatory agencies. In particular, performance measures are more acceptable to interested parties when they look valid and

fair, that is, when they have face validity. Such superficial appearances of fairness and relevance may be particularly important when the measures are used on a routine administrative basis, such as for making salary or promotion decisions, rather than for validating a predictor battery.

Any measure that is objectively scored may have an automatic edge in acceptability over measures that require ratings of some sort, because ratings frequently raise fears of rater bias or incompetence. Also, the more faithfully a criterion measure mimics the tasks one can observe workers performing on the job, the more job-related it will appear to be and thus the more readily accepted it is apt to be. Also, performance measures that show substantial mean group differences in test scores (e.g., by race or sex) are immediately suspect in the eyes of many interested parties. Measures that happen to be less face valid or to show larger group differences may in fact have equal or higher validity than measures that look more job-related or on which all social groups score equally well, but more supporting evidence is required to make the former equally defensible socially and legally.

Perceptions among interested parties of what constitutes the most valid and fair criterion measure may not agree with each other or with psychometric evidence—as has been the experience with intelligence tests in recent years. Nonetheless, these perceptions, whether accurate or not, still must be taken quite seriously because they can have great impact on the functioning of the organization.

Weighting Facets of Nonequivalence by Importance

Selecting a criterion measure from among alternatives involves two distinct processes: determining what the differences are among the measures and assigning utilities to each of those differences. The first is a matter of cataloging and measuring the sorts of differences just reviewed. The second process is one of weighting the differences by importance. In many cases trade-offs will have to be considered. One measure of job performance may be more expensive than another, but it may also be a more valid measure for the intended purpose. Some of the nonequivalencies can readily be expressed in terms of a common yardstick for measuring utilities—dollars, for example—but most will not be. Progress has been made in expressing differences in job performance in dollar terms (e.g., Hunter and Schmidt, 1983) and it is conceivable that all the nonequivalencies might be expressed in dollars, but it seems unlikely at this time. Reduction to a dollar metric is probably also unnecessary if the nonequivalencies can at least be rated by criticality or importance to the organization. Sinden and Worrell (1979) discuss various strategies for assigning relative values to "unpriced" goods for purposes of decision making.

As already discussed, assigning utilities to the different nonequivalencies

and even determining what they are depends on just what the goals of the organization are for performance measurement. Therefore, making a good choice depends on the clarity of the organization's goals. Psychometric standards are required for assessing nonequivalencies among job performance measures, but the choice among measures is ultimately a matter of economic judgment and social values: What kinds of performance does the organization want to obtain and reward? What is the organization able and willing—or unwilling—to "pay" to measure and obtain such performances? The bottom line is that a measure has to have marginal utility: the benefits flowing from the adoption of the performance measure must outweigh the costs that it imposes. Two measures are substitutable for a given purpose when their estimated utilities are the same and when those estimates are made with equal confidence, even though many particular facets of those measures may differ.

MAJOR ISSUES IN THE VALIDATION OF CRITERION MEASURES

The Nature of Validation for Criterion Measures

Much has been written about the meaning of validity and the forms it takes, such as construct, content, and predictive validity. The following sorts of issues have been debated, although most often in the context of predictor validation. Are there really different types of validity or are there only different validation strategies? Is content validity an aspect of construct validity, or might it be a form of test construction rather than of test validation? To what extent should one's validation strategy depend on the nature of what is being measured and on the purpose of measurement?

Lest one be tempted to dismiss the foregoing questions as merely semantic disputes of no import, it should be noted that very practical issues hinge on their resolution. Recommendations to adopt one performance measure rather than another often are influenced by beliefs about the kinds of validity that are preferable or sufficient, and court cases regarding personnel selection tests have been won or lost because of successful claims that one particular strategy should or should not have been used to validate them (Landy, 1986). In light of both the confusion regarding these issues and their practical import, any discussion of criterion equivalence must meet them head on and at least make clear the author's own stance toward validation.

The *validity of a measure* is a shorthand phrase referring to the inferences that may be drawn from the scores on that measure (Cronbach, 1971; Messick, 1975; Tenopyr, 1977). It follows, then, that validation is a process of hypothesis testing (Cronbach, 1971; Messick, 1975; Guion, 1976, 1978;

Landy, 1986). We may wish to draw a variety of inferences from a job performance test, depending on our purposes for using the measure—hence the frequent statement that a test has as many validities as it has uses.

Construct Validity and Relevance

Figure 1 helps to illustrate both the process of criterion development and the inferences we usually wish to draw regarding a criterion measure. This figure distinguishes between empirical measures of job performance (say, a work sample test) and the theoretical constructs those measures are presumed to operationalize (say, technical proficiency). Figure 1 also distinguishes constructs for individual-level job performance and constructs for organizational effectiveness.

These latter two types of constructs guide the development of job performance criterion measures. The organizational effectiveness construct represents the mission the organization wishes to accomplish by developing a measure of job performance; it is referred to here simply as the organizational goal. This goal could be one or more of any number of specific effectiveness goals, such as greater equity in personnel selection, higher production levels, improved product quality, increased military preparedness in one of the Services, or greater trainability or stability of the workforce. Setting such goals is beyond the scope of this paper, but it should be apparent from the foregoing list that setting such goals involves a careful consideration of the organization's needs, values, and priorities (Guion, 1976:793).

This organizational goal guides the search for the second construct—job performance. Choice of the performance construct, or conceptual criterion as it is sometimes called (Astin, 1964), is based on the researcher's or the organization's theory of what kinds of job performance will help the organization fulfill its stated goal; that is, a performance construct is selected on the basis of hypotheses about the value or relevance of different kinds of job performance to the organization (Staw, 1983). Often these constructs are not so much chosen as "negotiated" (Landy et al., 1983:1), because it is seldom clear just what kinds of performance are most likely to further the organization's goals. Identification of a performance construct, or conceptual criterion, for the jobs in question leads to the search for, or development of, one or more empirical measures to operationalize that construct. In some cases it is not feasible to operationalize the conceptual criterion, so a second-best substitute must be sought. Performance in combat is one example of a conceptual criterion for which a substitute performance construct must usually be found (Vineberg and Joyner, 1983).

Selecting and deciding how to operationalize a conceptual criterion involves clarifying which of the following aspects of performance is likely to

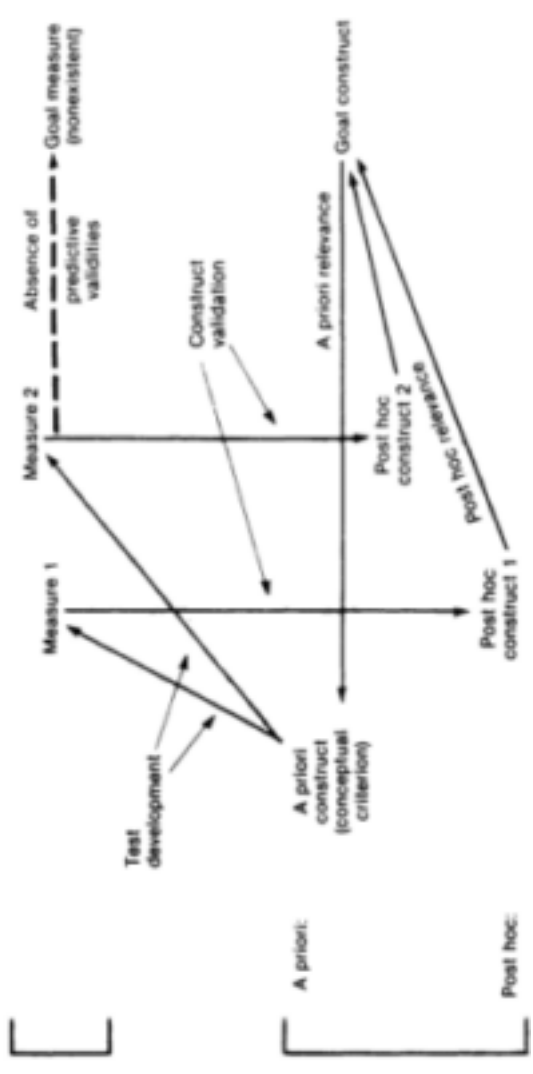


Figure 1 A schema summarizing the criterion development and validation process

be most critical to the organization in question. This list is illustrative, not exhaustive (see also Guion, 1976:793):

(1) maximal ("can do") versus typical ("does do") performance;
(2) performance in stable versus changing or disrupted environments;
(3) performance in well-defined versus ambiguous situations;
(4) initiative and innovation versus adherence to stipulated procedures;
(5) suitability only for the job in question versus for promotion or lateral transfer;
(6) performance on tasks performed as an individual versus (or including) tasks performed as a team;
(7) technical proficiency versus (or including) interpersonal effectiveness; and
(8) average performance level, consistency of performance, or proportion of work that is performed below acceptable limits.

These considerations affect not only the content and format of a criterion measure, but also how it should be administered and scored.

The general point is that all aspects of a criterion measure, from content to scoring, depend on the job demands that are identified as most important and whose performance is to be operationalized. Because jobs differ systematically in their major demands (e.g., Gottfredson, 1984), it can be expected that different kinds of criterion measures will sometimes be required for different classes of jobs. For example, relative to technical proficiency, interpersonal effectiveness is probably more relevant to organizational effectiveness in managerial and social service work than it is in clerical or crafts work. Work samples are not well suited to assessing interpersonal effectiveness, so we would expect ratings to be used more often in people-oriented than things-oriented work. Perhaps this is what is really meant sometimes by the term "method variance"—that different test types and formats are best suited for measuring different dimensions of performance; see Vineberg and Joyner (1983) for a thoughtful discussion of this point.

The criterion development sequence is illustrated in Figure 1 by the arrow from the a priori construct representing the organizational goal to the a priori job performance construct to be operationalized, and by the arrow from this performance criterion to the two different empirical measures that have been developed, in this illustration, to operationalize the job performance criterion.

Validation of a criterion measure involves testing the inferences underlying this development sequence, and it consists of two distinct steps: assessing the construct validity and the post hoc relevance of the criterion measure. These two kinds of inferences have also been referred to, respectively, as validity of measurement or psychometric validity and as validity of use of

the measurement or validity of propositions (Guion, 1983). The frequent failure to distinguish clearly between these two validation activities is surely a major source of the confusion in on-going discussions of validity and validation (cf. Guion, 1983:21).

Assessing construct validity is the process of determining to what extent the measure successfully operationalizes the conceptual criterion. Because it can be presumed that the operationalization of a conceptual criterion will be only partly successful, this step becomes one of determining what kinds of performance are actually being measured by the criterion measure, that is, of interpreting or attaching meaning to consistencies in scored responses to the test. These interpretations, or post hoc performance constructs, are shown for the two criterion measures in Figure 1. The conceptual criterion is often vague to begin with, so construct validation can be usefully described as a process of figuring out what components of performance have and have not been operationalized by the instrument, with the a priori conceptual criterion being only one guide to interpretation, and of then clarifying one's conceptual criterion in light of this knowledge.

The converse of construct validity is measurement bias, which refers to inappropriate inferences about what performances are actually being measured. Two generic sources of bias are contamination, which is the measurement of something that should not be measured, and deficiency, which is the failure to measure some desired aspect of performance. Two criterion measures may be equally construct valid (or biased) overall but have different contaminants or deficiencies. Depending on the projected use of the measure, any particular bias may or may not be a serious problem. If one's purpose is to validate a predictor battery, then bias in the criterion that is uncorrelated with the predictors will not affect the selection of the predictor battery, whereas predictor-correlated bias will adversely affect the selection of a battery and the weighting of its components (Brogden and Taylor, 1950). Of those biases that do adversely affect either personnel selection or performance appraisal procedures, some may have more serious consequences than others for the organization. The practical problem, of course, is that it is difficult to know whether or not a measure's biases are predictor-correlated, or if there even are any substantial biases.

The "criterion problem" arises, not because of the difficulties inherent in construct validation, but primarily from the need to assess the organizational relevance of a criterion measure or, more precisely, the relevance of the post hoc performance construct being measured. The relevance of a job performance criterion measure is its hypothetical predictive validity for predicting organizational effectiveness (cf. Nagle, 1953). Measures of organizational effectiveness may some day be available for computing predictive validities, but in their absence we must settle for judgments about relevance based on our theories about job performance and its impact.

Several other important points are apparent. One is that the actual (post hoc) relevance of a measure may differ considerably from what it was expected to be. One reason for this, of course, is the failure to successfully operationalize the original conceptual criterion. But another reason is that it may be decided that either the valid or the bias components of the measure have adverse consequences for the organization not previously considered, with the result that the organizational goals for performance measurement may be reexamined and modified. Another point is that inferences about criterion relevance are separate from but dependent on inferences about construct validity. If inferences about the construct validity of a measure change, inferences about the measure's relevance must also be reevaluated. Likewise, its relevance should also be reevaluated if the organization's goals for measurement change. Finally, I would argue that the ultimate concern in validating a job performance criterion for a particular use is to establish its relevance for that use. Determining construct validity is a means to that end; knowledge about the meaning of the performance being measured is a necessary but not sufficient element of the implicit or explicit theory justifying the adoption of the criterion measure.

This argument points to one difference between the validation of predictors and the validation of criteria that must be appreciated to avoid confusion when applying discussions of the former to the latter. If our purpose is only to predict with a measure, and we are able to compute a predictive validity, then we need *not* be as concerned with demonstrating the construct validity of the predictor measure (Tenopyr, 1977:49). The point is not that knowing a measure's construct validity (its meaning) is not incrementally useful beyond knowing its ability to predict some desired outcome, which is not true (Messick, 1975:956, 962; Guion, 1976:802), but only that the availability of predictive validities allows one to get some idea of a measure's relevance without first establishing its construct validity. When predictive validities are not available, as has been the case when validating job performance criterion measures, construct validity is absolutely essential to establishing the utility of such measures.

The Role of Content-Oriented Test Development

Claims for the validity of a particular test are often based on appeals to *content validity*, which refers to the instrument being comprised of items or tasks that constitute a representative sample of tasks from the relevant universe of situations (Cronbach, 1971). However, it has been argued persuasively that content validity is not a type of validity at all. For example, Messick (1975:960) argued that content validity "is focused upon *test forms* rather that test *scores*, upon *instruments* rather than *measurements"* [emphasis in original]. But inferences "are made from scores, and scores are a

function of subject responses. Any concept of validity of measurement must include reference to empirical consistency."

Following Messick, Guion (1978, 1983) and Tenopyr (1977) have also argued that it is more appropriate to refer to content validity as content-oriented test development or as content sampling, rather than as a type of validity, so that the psychometric concept of validity is not distorted. Content-oriented test construction strategies can contribute to valid measurement. For example, Messick (1975) described how controls can be built into a measure to preclude some of the plausible rival interpretations of scores on that measure. However, content-oriented test construction strategies seldom if ever are sufficient for demonstrating the construct validity of measures so constructed.

Appeals to content validity are nevertheless frequently made in an effort to demonstrate the validity of a criterion measure. Moreover, such appeals can short-circuit interests in doing empirical research on the meaning of the scores themselves, which is the essence of construct validation. For both these reasons, it is useful to look in some detail at the role of content-oriented test construction strategies in the validation process.

Referring again to Figure 1 helps to clarify the role of content-oriented strategies. Content validity is actually a test construction strategy in which a systematic effort is made to establish strong a priori presumptions of construct validity and relevance. Verifying the appropriateness of these inferences with empirical research using the measure goes beyond test development per se, and so goes beyond the notion of content validation. To provide strong a priori presumptions of construct validity for the scores obtained on a measure, content-oriented test development must carefully develop and document all of the following:

(1) a clear and explicit definition of the content domain;
(2) methods used to construct a sample of tasks from the content domain;
(3) methods used to develop test items for the content sample;
(4) test administration procedures and setting; and
(5) scoring methods.

Guion (1978) and Tenopyr (1977) have argued further that presumptions for construct validity on the basis of test construction alone are strong only when the content domain, (1) above, consists of simple, readily observable behaviors with generally accepted meanings. Note that this restriction probably rules out content valid tests for many jobs, in particular, for jobs requiring tasks that take a long time to complete, considerable mental activity (e.g., decision making, planning), or interpersonal or group activity.

Most claims for content validity in job performance measurement seem

to be based primarily on (1) and (2) above, and occasionally (3) as well. Claims based on the high fidelity of work samples would also seem to include aspects of (4), because fidelity refers to the realistic nature of the test setting and cues for performance as well as to the realistic nature of the test items themselves (Vineberg and Joyner, 1983). Despite the obvious importance of (1) through (3), (4) and (5) above are also essential, because appropriate inferences from test scores can depend heavily on the ways in which the test items are administered (e.g., test format) and scored (Guion, 1978; Vineberg and Joyner, 1983). For instance, although tests are routinely scored for level rather than for consistency of performance (Schoenfeldt, 1982), this choice of scoring method has no necessary relation to the content of a test. That choice does, however, have ramifications for the measure's meaning and relevance to particular goals. Moreover, the vast amount of evidence on the performance rating process and its susceptibility to bias (Landy and Farr, 1983; Landy et al., 1983) should, by itself, raise concerns about the appropriateness of claims for construct validity on the basis of content sampling whenever raters are needed to observe and rate performance—as they are in many work sample tests. For example, Pickering and Anderson (1976, as cited in Vineberg and Joyner, 1983) reported that military job experts or instructors typically fail to maintain standardized procedures when administering hands-on tests, often because they coach and give feedback as if they were training. And to take an example from the predictor domain, mental tests came under intense fire not only because of claims that their content might be culturally biased, but also because their stimulus conditions and scoring procedures might be less favorable to certain populations. Much empirical research was required to show convincingly that these plausible a priori claims were unjustified (Jensen, 1980; Wigdor and Garner, 1982; Gordon, 1987). A priori hypotheses regarding construct validity that are based primarily on content validity are stronger, then, to the extent that the measure looks like or mimics the job itself in all respects, from the tasks done to how task performance is evaluated.

The self-evident meaning of the content domain in a content-oriented measure, the great amount of care taken in enumerating and sampling tasks in that content domain, and the common practice of having persons familiar with the job and the organization rate the importance of tasks all create an aura, not only of construct validity, but of relevance too. While it might be agreed that the foregoing aspects of content-oriented test construction might improve construct validity, even though they cannot ensure or demonstrate it, such aspects of criterion development afford the resulting measures no special a priori claims to criterion relevance. To claim that more readily observable behaviors are more relevant than are increasingly abstract constructs of performance is to make a claim for the superiority of behaviorism over more cognitive theories of performance, which is something fewer

researchers have been willing to do in recent years. And care in sampling from a domain says nothing about the relevance of that domain.

Likewise, we may have no reason to question the judgment of subject matter experts when they rate the criticality of tasks in an effort to improve the relevance of a criterion measure. Nevertheless, no matter how familiar those experts are with the job and the organization, content-sampling strategies generally require those experts to work within the confines of the content domain defined by the researcher, which in turn is shaped by the researcher's own theories of work and organizations. At present, these theories seem to be largely implicit in content-oriented test development efforts. Although these implicit theories seem to be widely shared, or at least remain undisputed, they deserve greater scrutiny. The following look at the process of defining and sampling from a content domain, which is the centerpiece of content-oriented test construction strategies, illustrates that the construction, meaning, and relevance of criterion measures developed with such strategies remain as much a function of one's implicit or explicit theories about work as they do for performance measures developed in other ways.

Claims for content validity are most convincing when the content domain has been defined via a systematic analysis of the job independent of the people filling those jobs. The recommended procedure is usually to delineate the various discrete tasks performed on a job and then to determine both their criticality and the frequency of their performance. Tasks are then sampled for a criterion measure according to some combination of their frequency and criticality.

Traditional task analysis methods appear to conceptualize jobs as being built up of tasks whose demands do not vary according to the constellation of tasks in which they are embedded. Task-based criterion measures (whether they be work samples, paper-and-pencil job knowledge tests, or ratings) are thus composed of tasks that have been pulled out and isolated from the usual matrix of activity in a job. However, tasks pulled out of their usual context may present a partial or distorted view of a job's demands. This flaw may be similar to what Osborn (1983:8) has referred to as losing part of the content of a job in the "seams" of a task analysis. Workers often have to juggle tasks and set priorities for their performance (which is a task in itself) and to interrupt and restart tasks. It has been shown in other contexts that the intellectual difficulty level of a task can increase if it has to be performed simultaneously with another task (Jensen, 1987), but this sort of time sharing activity does not appear to be built into task-based performance measures (although it could be). Neither has the need to deal with the mistakes and incompetence of fellow workers been built into such measures, especially when jobs are interdependent, or to work under the distractions and other less-than-ideal conditions that characterize some jobs. Working

under stress, which is more typical for some jobs than for others, may also increase the difficulty level of many of the tasks of a job if it induces cognitive overload.

The variety of tasks performed may also increase the overall difficulty level of a job above that which would be expected from the sum of the difficulties of the individual tasks, even when they are performed serially. This hypothesis seems consistent with research (Christal, 1974) showing that the difficulty level of a job (which largely means the intellectual difficulty of the job) is partly a function of number of tasks performed as well as of the average difficulty level of the individual tasks comprising the job. The variety of tasks in a job may represent not only breadth of knowledge required but also a different and perhaps more important dimension of job difficulty—the infrequency or unpredictability of tasks performed. Strategies for sampling from a content domain often focus on tasks that are both critical and performed with some minimum frequency. The least frequent tasks are sometimes excluded from the content domain itself, even before their criticality is assessed. By excluding infrequent tasks, this strategy probably biases the sample of tasks toward typical, standardized, expected, and overlearned tasks. Such tasks are indeed important for organizational effectiveness, but to the extent that the proportion of the most critical tasks of a job are infrequent or unpredictable, the less the job can be standardized, the less the behaviors practiced, and the less often job aids produced to simplify the tasks. It also means that the job will require more continual learning and the exercise of more "judgment."

Cognitive abilities are somewhat more important in learning new tasks than in performing them after they are learned, at least in fairly simple jobs (Fleishman, 1975). Moreover, job demands for continual learning on the job and for judgment and acting under pressure are associated with higher intelligence requirements (Gottfredson, 1984). It might also be expected that unstable or changing organizational environments increase the unpredictability and novelty of tasks performed, which thereby increases the cognitive demands of the affected jobs. For example, the disruptions caused by military combat (e.g., lack of spare parts, damage to equipment, disrupted communications, and inadequate transport) all require improvisation and ingenuity, and the disruption may be especially acute for some occupational specialties (e.g., infantryman or tank crewman versus personnel clerk or automotive mechanic). Curran (1983) discussed the constant difficulty the military faces, for example, in developing task-based hands-on measures that measure coping with unanticipated problems in a job as well as with other demands in combat, such as the stress of personal danger, that are difficult or dangerous to include in a criterion measure.

In other words, the proportion of a job that consists of infrequent or unpredictable tasks is an important attribute of a job. If work content samples

capture only the stable and predictable components of a job, then they will lead to criterion measures that provide progressively less adequate representation of jobs with larger unpredictable components. High-level and more intellectually demanding jobs are less routinized, so it might be expected that traditional task analysis procedures provide a poorer representation of the content of such jobs than of less complex jobs.

The foregoing discussion illustrates that it is by no means an atheoretical task to define the content domain of a job or to sample from it. Those illustrations focused on the possible deficiencies of traditional task analysis methods for capturing the most important distinction among jobs in industrialized societies—general intellectual difficulty or complexity level of work performed (Gottfredson, 1985)—but the same examination could be extended to other dimensions of criterion performance and to other techniques for identifying a content domain. But these illustrations suffice to reinforce the argument that the construct validity and relevance of any criterion measure is established, not by detailing the techniques used to construct it, but by (1) research on the resulting test scores and (2) the adequacy of the theories of job performance and organizational effectiveness guiding the development and interpretation of the criterion scores and their relevance.

A great strength of content-oriented test construction for validation purposes, and a strength which I do not mean to minimize, is that it is a rich source of a priori hypotheses that can be empirically tested in validation research. As often noted, a clear specification of test construction procedures can serve as a good source of ideas about what the biases of a measure might be, and inferences about the meaning of criterion performances are supported to the extent that they survive plausible competing or disconfirmatory hypotheses about the meaning of those test scores (Gulliksen, 1968; Guion, 1978). The more good hypotheses about a criterion measure that are generated and tested, the more evidence there will be about its construct validity.

Criterion Bias against Subgroups

Concerns about test fairness in recent years have had a dramatic impact on the development, validation, and use of tests, and these concerns are a continual stimulant to regulation and litigation concerning personnel policies (Tenopyr, 1985). Now that evidence has accumulated that selection tests predict job performance equally well for blacks and Hispanics as for whites (Hunter et al., 1984), more concern has arisen that the criteria themselves may be biased. In view of this concern, it is important to address the issue of criterion bias against subgroups in the population.

Guion (1976:815) has remarked that "if the problem of investigating possible predictor bias is difficult, the problem of criterion bias is appalling."

One component of criterion bias has received extensive attention—potential rater bias against population subgroups such as women or blacks (e.g., Arvey, 1979; Landy and Farr, 1983:Ch. 5). This type of bias is a potential problem whenever examinee scores are assigned by raters or examiners. I shall focus here on what may be perceived as a more difficult issue. Whenever objectively scored tests show true mean subgroup differences in performance, might those tests be biased against the lower-scoring subgroup? For example, if racial differences are larger on job knowledge tests than on work sample tests, when both are scored in an unbiased manner, can it be assumed that either is biased against the lower-scoring racial group or that the former is more biased? And is it appropriate to adopt the performance measure with the smallest mean group difference if both tests seem content valid, as has sometimes been implied (Schmidt et al., 1977)?

Assessing bias against subgroups is an element of the larger process of determining the construct validity and relevance of a criterion measure. Previous investigations into the issue have focused on construct validity, that is, on questions of whether a measure really taps the performances it is presumed to tap and whether it does so equally well for all subgroups in question. However, bias against subgroups can also occur because of low relevance. Specifically, such bias occurs when (1) a criterion measure is either deficient or contaminated relative to the specified organizational goal (i.e., is not perfectly relevant) *and* (2) subgroups differ on the performance dimensions constituting the deficiency or contamination. For example, if a test (say, a job knowledge test) requires intellectual performance that is not required on the job, then it is biased against subgroups with lower average levels of the intellectual skills in question. Conversely, if a test (say, a work sample test) fails to tap intellectual performance skills required on the job, then the test is biased against the subgroup with the higher average levels of the skills in question. Any test that either over-or underweights the relevant dimensions of criterion performance will be biased against subgroups if subgroups differ on those same dimensions. Underweighting results in bias against the higher scoring subgroup and overweighting results in bias against the lower scoring subgroup.

In short, the most relevant criterion is the least biased against subgroups, because it rates people most closely in accordance with their scores on the performance dimensions valued most highly by the organization (cf. Cronbach, 1971, on the injustices introduced by test impurities and biased weights). When criterion measures differ in factor structure but are deemed equally but less than perfectly relevant (that is, when they have different contaminants or deficiencies), then both may be equally biased but against different subgroups. If race or sex subgroups do not differ on the underlying dimensions of performance being measured by a criterion measure, then that criterion measure will not be biased against any of those race or sex subgroups

even when it is less than perfectly relevant. However, that measure will always be biased against some people; in particular, it will be biased against people who score high on performance dimensions that are underweighted and against people who score low on dimensions that are overweighted.

If mean subgroup differences are larger on one criterion measure than on another, and if we presume that scoring procedures were unbiased in both, then the two measures are to some extent measuring different performance constructs. Thus, it cannot be presumed that the two measures are equally relevant to the organization. It is highly unlikely that two criterion measures that differ substantially in adverse impact (mean subgroup differences favoring the majority group) are equally relevant, even when they were designed to be so. It follows, then, that it is unwise to adopt the one with less adverse impact without evaluating the construct validity and relevance of both. Investigations into this issue using item response theory (Ironson et al., 1982) support this conclusion.

The fairness and appropriateness of the organization's goals against which relevance is assessed can be debated, as they often are, but that is not a psychometric issue (Gottfredson, 1986).

STRATEGIES FOR ASSESSING NONEQUIVALENCIES IN CRITERION VALIDITY

Assessing equivalence among alternative criterion measures is not a matter of computing some single coefficient of similarity. Instead, it requires the same ingenuity, research, and theorizing that are necessary for establishing the construct validity and relevance of any single measure.

Because criterion validation is a "prescription for hard investigative work" (Guion, 1976:777), it may be a tempting economy for an organization to limit in-depth assessments of criterion validity to only a single benchmark against which all others can be compared. However, such an organization will have difficulty knowing which alternatives to the benchmark are the more relevant ones if none is highly correlated with the benchmark. Two alternatives that are equally but not highly correlated with a valid benchmark may have different kinds of biases and therefore have very different prospects for furthering organizational goals. Riskier yet is the comparison of alternatives with a benchmark that is only presumed to be acceptably valid but with which no validation research has actually been conducted, as the best alternative may *not* be the one that is most similar to a flawed benchmark. If the organization has the resources to collect data for each of the alternatives under consideration, then relying on a priori judgments about validity or limiting validation efforts to a small proportion of the alternatives may be false economy.

Assessing Nonequivalencies in Construct Validity: Correlational Methods

I begin with the presumption that no two job performance measures (except parallel forms) measure exactly the same thing, even when designed to do so, and that the objective is to document both similarities and differences in the dimensions of performance tapped by two or more criterion measures. We know enough about current putative alternatives, such as job knowledge tests and work sample tests, to suspect that they do not measure exactly the same dimensions of performance in many jobs, even when they were all designed with the same general conceptual criterion in mind.

A factorial conceptualization of criterion performances is useful in discussions of criterion validation and equivalence. Whether a unidimensional or a factorially complex criterion measure is most appropriate for one's purposes and whether or not one is successful in developing a measure with the desired factor structure, any criterion measure can be conceptualized in terms of its factor structure—that is, as a weighted sum of different underlying dimensions of performance. Specifically, any criterion, Y, can be represented as the following sum

$$Y = a_1F_1 + a_2F_2 + \ldots + a_iF_i + e\ ,$$

where the F_i are the factors underlying the performance and the a_i are the weights for those factors in the criterion measure. Measurement error is represented by *e*, and the true score is represented by the sum of the remaining terms. Performance dimensions are unlikely to be uncorrelated in real life, but orthogonal factors are a convenient simplification for present purposes.

No criterion measure can be presumed unidimensional a priori, and many times we actually expect or want job performance measures to reflect performance on different and not necessarily highly correlated aspects of performance, all of which are of value to the organization (e.g., speed and quality of work). Univocal or unidimensional criterion measures are simply those that have nonzero weights on only one performance factor. Measures that are equivalent in true factor structure tap the same underlying dimensions of performance and weight them the same.

Wherry et al. (1956) provided a useful framework for exploring factorial equivalence (see also Gulliksen, 1968, and Smith et al., 1969, for other discussions of equivalence). Wherry et al. investigated seven basic proposals for computing estimates of overall degree of criterion equivalence. The critical analysis of these indices, which is presented below and draws heavily from the Wherry et al. paper, shows that estimates of overall degree of equivalence are seldom sufficient information for assessing the relative validity of two measures, they often differ widely from one index to another,

and they can be very misleading. This critique is provided below partly to reduce temptations to unnecessarily limit analyses of equivalence to the computation of similarity coefficients. A discussion of the different indices also is useful because it reveals correlational methods for investigating the nature of equivalencies and nonequivalencies among criterion measures, and thus of determining the proper interpretation of alternative criterion measures. To some extent, the following analytic strategies constitute guides to thinking about criterion equivalence more than they do methods of empirically investigating it, because sufficient data will not always be available to utilize them.

Wherry et al. examined variations of the following seven indices created by varying the interpretation of "similarity of profile" to measure (a) level, (b) shape, or (c) a combination of shape and level:

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

Wherry et al. computed and compared all of these alternative indices of equivalence using job performance data they had collected for the military, and they compared the measures in terms of factor theory. They also intercorrelated estimates of equivalence generated by the different indices and factor analyzed those correlations to discern the major differences among the different indices of equivalence. Although the indices of equivalence often produced estimates that were at least moderately correlated, no two led to exactly the same conclusions about level of equivalence among their criterion measures and some led to quite different conclusions. Wherry et al. concluded that the measures of similarity in profile shape were the most appropriate, overall, so measures involving profile level will be ignored in this paper. Moreover, the measures involving profile shape are sufficient to make the point that assessments of criterion equivalence require a validation

process rather than the computation of a simple coefficient of equivalence.

The discussion begins by assuming ideal measurement conditions, including perfectly reliable criterion measures and a very large and representative sample of the population to which generalizations are drawn and in which each person has scores available on all relevant variables. The effects of these measurement limitations on estimates of equivalence and on the possibility of even assessing equivalence are discussed briefly at the conclusion of this paper. All indices are described before they are evaluated.

Figure 2 elaborates the Wherry et al. analyses by clarifying the substantive differences among the different indices and the severe limitations of most of them by putting those indices into a common broader perspective. The rows of Figure 2 represent data for each of the criterion measures (or their components) under consideration in a study. Seven types of data about the criterion measures are represented by matrices A through G; each of the matrices or certain combinations of them produce different indices of factorial equivalence. Attention will be restricted to measures of similarity of profile shape, which means that all the indices of equivalence are calculated by correlating the data for one criterion measure (in one row) with the data for the other criterion measures (in the other rows of the matrix in question).

The first six of Wherry et al.'s approaches to measuring criterion equivalence, in terms of similarity in profile shapes, correspond to Figure 2 as follows:

(1) matrix A
(2) matrices C, D, and E
(3) matrices D and E with individual criterion measures added by extension
(4) matrix B
(5) matrix G
(6) matrix F

One measure of equivalence not reviewed by Wherry et al., based on matrices C and D, will also be discussed. This paper does not discuss Wherry et al.'s seventh approach—cross-validation/criterion extension—because it is basically a composite measure of differences in the reliability of beta weights and in the predictability of two criterion measures from each other.

Matrix A consists of the scores of individual examinees on each of the criterion measures. The index of equivalence derived from this matrix is simply the zero-order correlations among the criterion measures (which are assumed for the moment to be perfectly reliable). A high correlation means that persons who score high (or low) on one measure score high (or low) on the other.

Matrix B represents scores (0/1 for absence versus presence) indicating

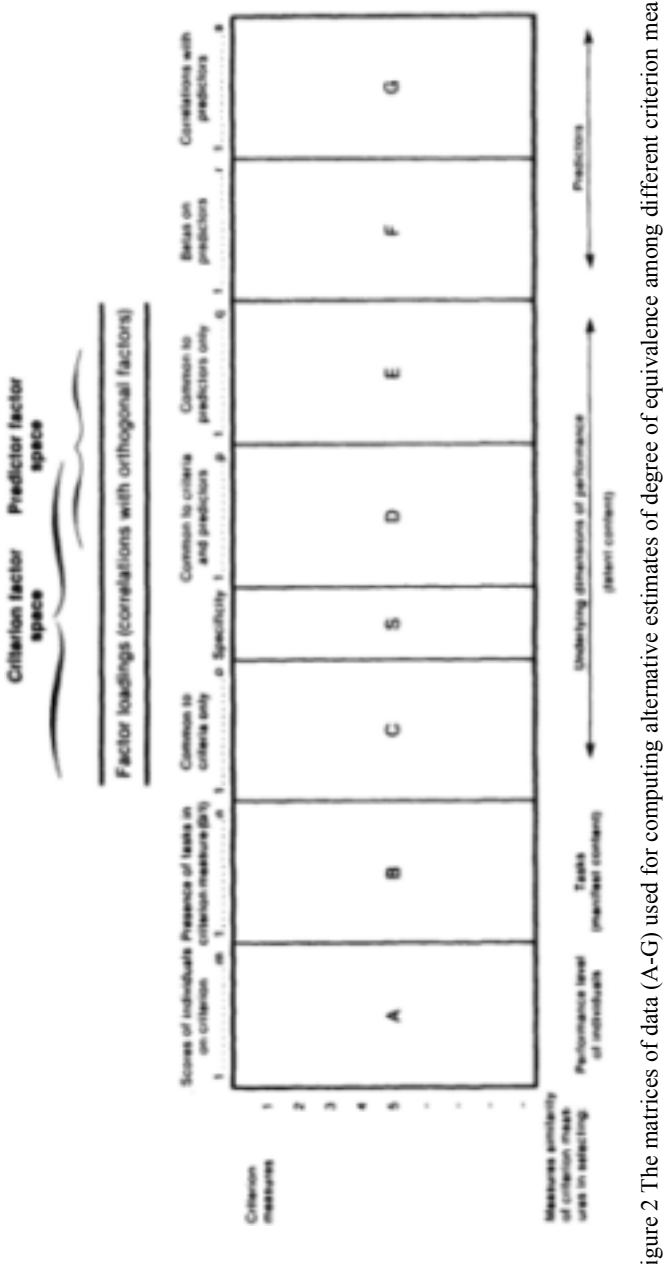


Figure 2 The matrices of data (A-G) used for computing alternative estimates of degree of equivalence among different criterion measures

which tasks from the total task domain are actually sampled in each of the criterion measures. The index of equivalence calculated from correlating the rows of this matrix reduces simply to a measure of task overlap for any two criterion measures (Wherry et al., 1956). The greater the degree of overlap, the more similar the manifest content (e.g., items) of the two criterion measures. (It should be noted that this measure relates to characteristics of the criterion measure, not to people's scores on that measure, and so provides no empirical evidence concerning construct validity.)

Matrices C through E represent factor loadings of the criterion measures on different underlying performance factors. Matrices C, D, and E are based on the very useful distinction Wherry et al. drew among three types of underlying performance factors, which for ease of discussion are assumed to be orthogonal: factors common to two or more criterion measures but not to any predictors (matrix C), factors that are common to at least one criterion measure and one predictor (matrix D), and factors found among the predictors only (matrix E). Matrix S represents the specificity of a criterion measure, that is, the reliable variance it does not share with any other variable in the analysis. The criterion factor space consists of matrices C, D, and S; the predictor factor space consists of matrices D and E. Thus, matrix D represents the overlap between the predictor and criterion factor spaces, and matrices C, D, E, and S represent the combined factor space represented by both predictors and criterion measures.

Three different indices of overall degree of equivalence can be conceptualized from different combinations of these four matrices (and actually computed when sufficient data are available)—one representing an analysis of the criterion space (matrices C and D), one a joint analysis of both the criterion and predictor spaces (matrices C, D, and E), and one an analysis of the predictor space (matrices D and E) with criterion measures added individually by extension. Matrix E does not actually affect potential computations of degree of equivalence in the second two analyses, because all criterion measures have zero loadings by definition on factors in this matrix. The first two analyses include matrix S implicitly, but the loadings in that matrix do not affect estimates of degree of equivalence because they are always zero by definition for all but one criterion measure, which means that cross products with those loadings are always zero. Although they do not affect computations of degree of equivalence, it is still important to attend to matrices E and S because they provide clues to the nature of equivalence and nonequivalence, as will be discussed later. The three indices of equivalence calculated from factor loadings represent the equivalence of criterion measures in, respectively, the criterion factor space, the joint criterion-predictor space, and the predictor factor space, where equivalence is defined in effect as having proportional weights on all factors. Although the first two methods are in a sense logically identical, it is shown below that the actual

estimates of overall criterion equivalence they would provide are not the same.

Wherry et al. referred to the third method as the criterion extension method. This method of analysis might be used when an investigator has the results of a factor analysis of the predictor measures only. Specifically, if correlations of the criterion measures with the predictors are also available, then the factor loadings of the criterion measures on the factors in the predictor space can be estimated. Also, if scores from different criterion measures are not all available from the same sample, and cannot be directly compared, an investigator might want to estimate the loadings of different criterion measures on a common or standard predictor factor space without including the criterion measures in the factor analysis, because including one or more criterion measures in the analysis might substantially change the factor solution and differentially so from one criterion to another.

Dotted lines are drawn between the four matrices of factor loadings to illustrate that any particular performance factor may be allocated to different matrices depending on the specific criterion and predictor measures that are included in the analysis. For example, whether a specific criterion factor falls into matrix C or matrix D depends entirely on whether a predictor tapping a factor in matrix D happens to be included in the analysis. Likewise, if we increase the number of criterion measures in the analysis, we are likely to cover more of the theoretical criterion factor space. In all likelihood, this will also reduce the specificity variance of most or all of the criterion measures. Depending on how much the predictor factor space overlaps the criterion factor space, adding predictor variables to the analysis can have the same effect of reducing specificity variance in the criterion measures. To the extent that new variables tap new sources of variance in the criterion or predictor factor spaces, the nature and number of factors appearing in a factor analysis can also be expected to change. These facts will be shown later to be extremely important.

If each criterion measure is in turn regressed on the same set of predictors (as when a predictor battery is being validated for each criterion measure from the same pool of predictors), the resulting prediction equations will consist of sets of beta (standardized regression) weights for the predictors. The rows of matrix F represent these beta weights for each criterion measure. A high estimate of overall equivalence using this method would mean that the same predictors are most useful in predicting the two sets of criterion scores and that the importance of the predictors relative to each other is the same (i.e., the regression weights are proportional).

Matrix G represents the zero-order correlations of the criterion measures with a set of predictor measures. That is, it represents a matrix of the validities of the predictors for predicting the criterion measures (or vice versa). A high estimate of equivalence with this index means that the pattern

of correlations of one criterion measure with a set of predictors is the same as the pattern of correlations of the second criterion measure with the same set of predictors; the correlations are not necessarily the same, but they are at least proportional. The term predictors is used here in order to distinguish clearly these noncriterion variables from the criterion measures, but there is no implication that the former are otherwise restricted in type. They may be measured concurrently or predictively, and they need not be candidates for inclusion in a personnel selection battery.

As noted in Figure 2, the easiest way to conceptualize the substantive differences among the different indices of criterion equivalence is to observe what their units of analysis are or what they "select" for: individuals' criterion performance levels (matrix A), tasks (matrix B), underlying factors of performance (the factor loadings in matrices C, D, E, and S), and predictor measures (matrices F and G). The indices of equivalence derived from factor loadings can be further subdivided into those that select for factors in the criterion space, the predictor space, or a combination of the two.

Under certain conditions, some of these different matrices will produce identical estimates of overall equivalence. For example, if predictors are uncorrelated with each other, beta weights and predictive validities will be identical, meaning that entries in matrices F and G will be the same. Under most conditions, however, the different matrices of data produce different estimates of equivalence—not only in absolute level of equivalence, but also in which criterion measures are most nearly equivalent to each other. Nonetheless, the analyses leading up to the computation of these indices are very useful in assessing the nature of criterion equivalencies and nonequivalencies and so in assessing the construct validity of each criterion measure. The strengths and limitations of the analyses associated with each matrix are discussed next.

A serious limitation of three of the indices stems from the fact that they are entirely predictor dependent, that is, they rely entirely on data about the relations of the individual criterion measures to a set of predictors and not at all on data about the direct relations of the criterion measures with each other. The three predictor-dependent measures are those that assess criterion similarities in loadings on the predictor factors (matrices D and E via the criterion extension method), in beta weights (matrix F), and in predictive validities (matrix G). (In the former case, criterion measures are not included in the factor analysis, so matrices D and E are indistinguishable and reduce to E alone, but probably with at least somewhat different factors.) The serious problem with predictor-dependent indices of equivalence is that they cannot register similarities across criterion measures that are not also shared by the available predictors. If two criterion measures share some common performance factors, this criterion overlap will be apparent only if predictors of these same factors are included in the analysis. For example, if

two criterion measures both tap performance on psychomotor tasks, their apparent degree of similarity will be higher when a relevant psychomotor ability test is included among the predictors than when one is not. Furthermore, the rank order of equivalence of one criterion measure with several others can change when the set of predictors is altered. For example, if one criterion taps both cognitive and psychomotor task requirements, if a second taps primarily cognitive task performance, and if a third taps primarily psychomotor tasks, then the first criterion measure will appear most nearly equivalent to the more cognitive criterion measure if the predictors are cognitive tests, but it will appear most nearly equivalent to the psychomotor criterion measure if the predictors are psychomotor tests. In fact, however, the first criterion measure may be equally correlated with both of the others. Predictor-dependent methods provide the clearest evidence regarding the factorial equivalence and construct validity of criterion measures when there are high multiple correlations between the predictors and each of the criterion measures.

At this point it is useful to note that the measure of equivalence based on predictive validities (matrix G) resembles a formalization of a commonly used technique in construct validation. If two measures have high correlations with the same variables and low correlations with the same variables, this is evidence that they measure the same theoretical construct (although it still may not be clear what that construct is). But the index of equivalence based on similarities in predictive validities will provide only a pale and sometimes misleading imitation of this construct validation strategy if the predictors are restricted to variables that are candidates for inclusion in a personnel selection predictor battery. Such predictors constitute only a subset of the variables of theoretical interest and exclude those known to have only negligible correlations with the criterion measures. If, in addition, the predictors are all moderately to highly correlated with each other, as would be the case with most mental test batteries, then there will be little systematic variability among the predictive validities with which to establish reliable profiles of validities. More useful information about relative construct validity is obtained by employing a diverse set of predictors, only some of which would ever be seriously considered as predictors for personnel selection purposes. To be most useful in construct validation research, the predictors should themselves have high construct validity and be embedded in a valid theory of human performance. The same predictors may be interpreted differently depending on one's theory about the organization of abilities and behavior, and these differences in the interpretation of predictors can lead to different interpretations of the criterion space. Thus, one's interpretations of the predictors should be carefully considered.

Another limitation of the predictor-dependent measures, and also of the direct comparisons of criterion measures via factor analysis, is that the

apparent degree of equivalence of two measures can change depending on the other variables that are included in the analysis, whether they be criterion or predictor measures. That is, some measures of overall equivalence can produce quite different estimates of equivalence for the same sets of criterion scores depending on the other types of data used in calculating those estimates. This problem of invariance plagues all of the indices in Figure 2 except zero-order correlations and task overlap (matrices A and B). The addition of new variables to a factor analysis can change the factor solution, which in turn can change the correlation of factor loadings across the different criterion measures. The less correlated the new measures are with the old, the more serious this problem is likely to be. To take another example, beta weights are very unstable under certain conditions. For example, the size of a beta weight for a predictor decreases with the addition to the regression analysis of other predictors highly correlated with that first predictor. Thus, if predictors are highly correlated, they cause problems for the beta weight method; if they are not highly correlated, they cause problems for the factor loading methods. Estimates from the factor loading methods are also sensitive to factor rotation, which further implies that the use of such methods requires a good theoretical rationale for the factor structure or rotation method chosen.

Even if we assume that the previously noted problems of invariance have been mitigated by settling on a theoretically sound solution to the factor analysis, there is still the question of whether similarity in factor loading profiles adequately operationalizes the notion of factorial equivalence. Similarity in shape of factor loading profiles can be highly correlated with the zero-order correlations between criterion measures (matrix A), as Wherry et al. found in their data, but high correlations need not occur. For example, the factor loadings .1, .2, and .3 are perfectly correlated with the factor loadings .2, .4, and .6, but the implied zero-order correlation between the two hypothetical criterion measures that they represent is only .28 (as calculated from the summed cross-products of the loadings, and assuming that the factors are orthogonal). In addition, the proportion of the variance in the first criterion that it shares with the other measures in the analysis (its communality) is only .14, whereas the communality of the second criterion measure is .56 (as calculated from the sum of squared factor loadings). If the analysis has been restricted to criterion measures only, these communalities suggest that the first criterion may have little in common with other measures of job performance. Such a large uniqueness can signal either an advantage or a disadvantage, so being aware of that degree of uniqueness and understanding its content can be important.

Although the factor loading indices are not appropriate for determining degree of factorial equivalence, factor analyses are very useful for investigating the nature of criterion equivalencies and nonequivalencies. Factor

analyzing criterion measures or their components can provide clues about what underlying performance dimensions the criterion measures have in common that may not be apparent from their manifest content. It can also provide clues to the nature of their nonoverlap. When performance factors in one criterion measure do not overlap the performance factors in another measure, then that nonoverlap constitutes either contamination in one criterion measure or deficiencies in the other if they have been designed to operationalize the same performance construct.

Factor analyzing the criterion measures together with predictor measures—or better yet, correlating theoretically sound predictor factors with criterion factors generated separately—can further illuminate the nature of the common and noncommon factors underlying criterion performances. Finally, joint analysis of criterion and predictor spaces will help reveal the amount and type of overlap of the criterion and predictor spaces themselves. Information about the degree and type of overlap of predictor and criterion spaces is not itself relevant to the selection of criterion measures beyond what it contributes to an understanding of those measures, but it relates to one of the fundamental problems in personnel selection, job classification, and validity generalization—the need for more knowledge about the links between the task requirements and the ability requirements of jobs (Dunnette, 1976). Such knowledge is valuable for developing predictor batteries and is ultimately necessary for a comprehensive theory of job performance, which itself might guide future criterion development.

Turning to one of the two remaining indices, task overlap does not seem to be a generally viable method for estimating degree of factorial equivalence. The very different nature of many alternative criterion measures, such as work sample tests composed of specific work tasks versus supervisor ratings of more general behavioral dimensions, makes it difficult if not impossible to assess their task overlap and thus to quantify criterion equivalence via this means. However, the pattern of correlations among the scores people obtain on different tasks may provide clues about why certain criterion measures share some underlying performance factors but not others, how particular criterion measures may be deficient or contaminated, and how the various elements of a criterion measure might be broken out to create subtests of the criterion measure. Those components from the various criterion measures might themselves be used in a factor analysis of criterion scores to gain a more detailed understanding of the criterion space, or the analysis might begin with them if there are too few criterion measures for a factor analysis of total test scores. They might even be considered potential building blocks for a new and better composite criterion measure.

The one remaining measure of equivalence—the (disattenuated) zero-order correlation between criterion measures—is the most appropriate index of degree of overall equivalence in factor structure. However, by itself it

provides only limited information for making decisions among criterion measures because it says nothing about the nature of the equivalencies and nonequivalencies. For example, two criterion measures may be highly correlated not because they tap the same desired performance measure, but rather because they are contaminated in the same way. Also, one measure can be equally correlated with two others, but for very different reasons. One may share a desired performance factor whereas the other may share only a contaminant. Furthermore, it may be possible to reduce contamination if it can be identified.

Criterion validation is hampered by a lack of knowledge about the organization of the job performance domain. Compared to our knowledge of the human abilities predictor domain, knowledge of the criterion domain is meager. It can be argued (Guion, 1976, 1985) that the first requirement, yet unmet, for establishing a systematic procedure for identifying promising criterion (or predictor) measures is the search for the fundamental constructs of job performance. Factor analytic methods have been stressed in this discussion, which accords with previous discussions of equivalence (Wherry et al., 1956; Gulliksen, 1968) and previous practice in investigating the criterion domain (e.g., Richards et al., 1965), but other methods may be equally or more useful. Emerging taxonomies of human performance (Fleishman and Quaintance, 1984), although yet of only limited applicability to criterion development, illustrate the variety of conceptualizations of the performance domain that are possible and that might be tested. Tenopyr (1977) has also described the value of a taxonomy of constructs in the context of discussing the development and validation of performance measures.

Assessing Nonequivalencies in Construct Validity: Other Methods

A good understanding is required of the internal psychometric properties of all measures being used, otherwise faulty inferences may be drawn about the construct validity of each and about the nature and degree of relation they have with each other. This is especially so when severe measurement limitations distort the correlations among variables in an analysis. If predictors are used to aid in the interpretation of criterion measures, then their properties should receive the same scrutiny.

Distributions of scores should be examined to check for outliers because outliers can have large effects on any parameters calculated. Item analyses can be performed to assess the discriminability of the test along different ranges of total test scores; for example, they might reveal ceiling or floor effects. It might be noted in this regard that personnel selection tests are often designed to discriminate best at certain ranges of performance; for

example, ASVAB subtests have been designed to discriminate better in the lower ranges of mental ability, except for the mathematical tests, which are more difficult (U.S. Department of Defense, 1984; Ree et al., 1982).

It may also be of interest to examine the manifest content of items ranked differently or similarly in difficulty level within a measure. Bivariate distributions of total test scores on two measures can provide additional insight into each criterion as well as into their relations with each other. For example, it might be found that people who score low on a job knowledge test also score low on a work sample test, but that there are large differences in the job knowledge scores of people who score high on the work sample test, as might happen if the work sample test fails to discriminate well among the better workers. Which criterion is to be preferred depends on one's particular goals for measurement, so it is important to know how such differences among the criterion measures relate to one's goals. If one wants to exclude poor performers, discrimination is not required in the higher ranges. It would be required, however, if one's purpose required the identification of high performers.

It may also be useful to look at bivariate distributions for particular subsets of items. If many of the tasks included in different task-based measures for a job are identical (e.g., in work sample tests and task-level ratings), then one would hope that responses to the items concerning a task in one criterion measure would be highly correlated (within the limits of reliability) to responses to items on the same task in the other criterion measure. If they are not, examining the patterns of responses and their relation to specific predictors or to factors in the predictor space might explain why such differences occur. Close attention to differences between the measures in how items were developed, administered, and scored for the task might also provide an explanation of such unexpected differences in performance on presumably the same tasks.

It might become apparent during these analyses that some of the criterion scores need to be transformed. For example, scores may typically be presented in percentiles for some tests but in standard scores for another. Using percentile scores may not cause much distortion in results, particularly because the correlation coefficient is not very sensitive to differences in scale units (Gorsuch, 1974:268), but using such scores is a potential complication that can easily be avoided. All the foregoing data on the internal properties of the criterion measures will also aid in the appropriate interpretation of any correlational analyses when measurement limitations such as unreliability or differential restriction in range on criterion performances are apt to distort the correlations.

I have focused here on only those data that are likely to be present during the first stages of the criterion validation process because organizations will begin selecting from among various potential measures at this stage of the

research. Other sorts of data, however, might be collected to investigate the construct validity of any particular measure or to document the exact nature of differences among several. For example, interpretations of the meaning of different measures can be tested by subjecting applicants or workers to experimental treatments (e.g., specific training programs) that would be presumed to change scores if the interpretation of the construct is valid (Cronbach, 1971:474; see Smith, 1985, for an example of this approach). Also, data on additional theoretically relevant variables might be collected to test emerging hypotheses about differences in construct validity and bias across criterion measures.

Assessing Nonequivalencies in Criterion Relevance

As discussed earlier, the relevance of a criterion can be conceptualized as its hypothetical predictive validities, where the predictions concern the impact of that performance on the fulfillment of organizational goals. It is theoretically possible, but seldom if ever feasible, to generate predictive validities empirically. In order to assess actual (post hoc) relevance, then, three things are required: (1) a clear specification of the organizational goals that the performance measure is intended to serve; (2) knowledge of what performance constructs the criterion measure actually measures (its construct validity); and (3) theory or evidence about the impact of the measured job performance on the organization, including the impact of both the contaminants and desired performances. The process of assessing relevance may actually function to clarify one or more of these elements, because goals, constructs, or theory may have been vague to begin with. The failure to clarify all three elements means that the organization risks not developing the most useful criterion that it might have otherwise.

Specifying organizational goals for measurement is beyond the scope of this paper, and the determination of construct validity has already been discussed. The importance of the third element—theory—is argued in the discussion of content and construct validity but is of more systematic focus here. By theory I mean well-reasoned and explicit hypotheses, whether they be based on practical experience in organizations or extracted from research on performance appraisal, personnel selection, organizational behavior, or other related topics.

To be persuasive, such hypotheses should specify the intervening mechanisms or processes by which individual-level performance has an impact on the functioning of the organization. The value of any particular dimension of performance can vary according to the organization's particular needs, goals, and structure, but the following examples illustrate the ways through which specific kinds of performance may affect organizational functioning.

(1) Worker error or inefficiency may increase down time for equipment, processes, or other workers (e.g., the failure to resupply or repair parts on time);
(2) Worker error or carelessness can result in costly damage to equipment or materials or in injury to self or others;
(3) Serious worker errors or inconsistency of performance can damage the organization's reputation;
(4) Poor or erratic performance can increase needs for supervision. It can also increase the aptitude demands among coworkers in interdependent jobs (e.g., to compensate for the poor performance of the worker in question); and
(5) Lack of technical competence in a supervisor can decrease performance and morale among subordinates.

As these examples suggest, the value of performing a task or job well stems from how and where the task or job is embedded in the work of the larger organization. These examples also suggest that evaluating and setting standards for performance in any one job, as is done partly by the choice of criterion measures for that job, should be done with an eye to the effects of that choice on performance standards in other jobs. For instance, underestimating the utility of certain dimensions of performance, or accepting what appear to be inconsequentially lower levels or consistency of performance in several jobs, could have the unanticipated consequence of drastically increasing requirements for supervision, which amounts in effect to raising the work demands of supervisory workers or increasing their number. This may or may not be the most effective use of the available manpower and at the very least, if not anticipated, could cause temporary disruption of organizational activities. This example also raises the issue that while criterion development was guided by specific organizational goals that may have been restricted in scope, evaluation of criterion measures must also be concerned with the possible unanticipated effects on other organizational goals. Uhlaner and Drucker (1980) and Staw and Oldham (1978) exemplify work in which individual-level performance is viewed from such a systems perspective.

The lack of comprehensive and integrated theories of job performance and of its relevance impede the evaluation of alternative performance criterion measures. However, the evaluation of alternative measures affords a great opportunity to further the development of such theory (cf. Vineberg and Joyner, 1983), particularly if it forces one to articulate and test a theory (or part of a theory) of job performance. This process of clarifying assumptions and hypotheses is often seen as a beneficial by-product of modeling (Campbell, 1983), which seems to be borne out by efforts to model job proficiency. The causal modeling work by Hunter (1983) and Schmidt et al.

(1985), that focused on the relations of different performance measures with each other and with various predictors of performance, encourages the explication of just how and why criterion measures differ in the constructs they measure, how they are causally related, and why their relations to each other and to various predictors may differ systematically as a function of organizational differences in training and job standardization. To take another example, Smith's (1985) work relating global and specific measures of job satisfaction helps to illuminate the breadth and magnitude of relevance that different types of criterion constructs may have.

Finally, the process of assessing criterion relevance can also be a process of improving criterion relevance. One need not choose from among the existing measures. If serious contamination or deficiency is discovered in even the most promising alternatives, then those criterion measures should be improved. If a clarified and more relevant conceptual criterion emerges during the validation process, then the original criterion measures might be further tailored to approximate this improved conceptual criterion. For example, if it is decided that the dimension of performance given the greatest weight by a criterion measure is less critical to the organization than is another dimension, then some reweighting of the criterion measure's components should be considered to give greater weight to the more critical performances.

Before leaving the issue of equivalencies in criterion relevance, it is important to clarify an issue that can lead to confusion. It could be argued that one need not be interested in how similar two types of criterion performances themselves are in relevance when the purpose of performance measurement is to develop a predictor battery for selecting and classifying workers. Rather, the argument goes, similarity of predicted rather than of actual performance levels is of more interest here, because applicants will be selected and classified on the basis of their predicted scores. Thus, even though the job performance factors tapped by one measure may be more relevant than those tapped by another, the two measures are nevertheless substitutable if the prediction equations they validate lead to the same decisions about applicants, such as the hiring of the same people. (As Schmidt, 1977, has noted, the prediction equations themselves need not be identical to produce essentially the same hiring decisions.)

Although this argument has merit, it refers not to the relevance of alternative criterion measures but to the relative utility of the predictor batteries developed in research with those criterion measures. It should be understood that similarities in the utility of predicted scores, despite differences in the relevance (and potential utility) of actual scores, may result from an unnecessarily restricted pool of predictor variables. Dissimilar criterion variables cannot be presumed to be fully predictable by the same predictor equations (but see arguments by Schmidt, 1977; Schmidt et al., 1981). For example, if

the most relevant criterion is multidimensional, then one should expect similarly multidimensional predictor batteries to best predict the criterion performances. To illustrate, say that a particular work sample test reveals physical as well as intellectual dimensions of performance in a given job, or that a peer rating system for a different job reveals interpersonal as well as intellectual dimensions of performance in that job. It would not be wise in either case to limit, unnecessarily, one's validation research to highly unidimensional predictor batteries, such as the ASVAB (Cronbach, 1979; Jensen, 1985), because one-factor batteries can predict only the same single dimension of performance across different criterion measures no matter how different those criterion measures are otherwise. None of the nonintellective components of the different criterion performances would be predictable from the cognitive battery alone. No matter how carefully developed or relevant those other components of the criterion measures might be, they would remain unexploited. It might not be possible to find or develop valid predictors for the various relevant nonintellective factors of performance (say, some aspects of interpersonal competence), thus leaving the criterion measure underutilized. Nonetheless, the relative utility, and thus the substitutability, of criterion measures should not be assessed until the dimensionality of the criterion performances has been investigated and the search for feasible, valid predictors has been exhausted.

The Impact of Measurement Limitations on Validation

The discussion of methods for assessing factorial equivalence among criterion measures assumed for convenience that there are no measurement limitations. Unfortunately, this is never the case and limitations are sometimes severe. Recent advances in meta-analysis have shown how interpretations of predictive validities have gone astray in the past because of the failure to appreciate fully the impact of measurement limitations (Schmidt et al., 1976; Schmidt and Hunter, 1981). Interpretations of data on criterion equivalence are no less vulnerable to the same limitations. Four measurement limitations are reviewed below.

Sampling Error

The smaller the sample size, the larger the sampling error and the weaker the inferences drawn from the research results. Therefore, a small validation study provides only weak evidence. Larger studies and more studies, if the latter are subjected to meta-analysis, can provide much stronger evidence. In their own meta-analyses of the predictive validity of cognitive tests in personnel selection, Schmidt and Hunter (1981) discovered that 75 percent of the variance in validity coefficients was due to four statistical artifacts

and that fully 85 percent of the variance due to artifacts was due to sampling error alone, indicating the importance of fully appreciating that particular measurement problem.

It follows then, that a small empirical study may do little to support or disconfirm one's a priori hypotheses about the construct validity of a particular criterion measure. Until a sizable body of criterion validation research accumulates, organizations seeking criterion measures should conduct as much validation research as feasible, conduct it as carefully as possible, and ascertain the statistical power of their proposed analyses before the research is actually conducted.

Unreliability

The less reliable a measure, the lower its observed correlations with other variables, all else equal. Even if two criterion measures have the same factor structure, they will be correlated only to the limit of their reliabilities. Thus, the least reliable measure will have the lowest observed correlations with the other criterion measures, all else equal. Estimating the true score correlation between two criterion measures requires that the observed correlation be divided by the product of the square roots of the reliabilities of the two criterion measures.

When the objective of an analysis is to understand the content of a criterion measure and its theoretical relations to the predictor or criterion factor spaces or to other variables, all correlations must be disattenuated by the relevant reliabilities. This includes the predictors. When predictors are validated against criterion measures for selecting a predictor battery, it is common practice to disattenuate the correlations between criterion and predictor measures for unreliability in the criterion but not for unreliability in the predictor. The reasoning is that we want to know how well the predictor can predict true criterion performance levels, but we can select individuals only according to their observed, fallible scores on the predictor. The situation is different when the aim is to understand the true relations among test scores, as is the case when trying to discover what dimensions of performance a criterion measure does and does not tap. For example, if the reliabilities of the predictors differ substantially, we cannot expect factor solutions that include the predictors to be the same when correlations have been corrected for unreliability in the predictors as when they have not.

As noted earlier, accurate reliabilities may be difficult to determine and under-and overcorrections can occur, but complete disattenuation should be attempted whenever possible for analyses exploring the nature of criterion equivalence. Analyses can be repeated with different estimates of reliability to determine how sensitive interpretations are to possible errors in estimating reliability.

Differential Restriction in Range across Criterion Measures

The concern here is not with restriction in range on the predictors, except indirectly, but with restriction in range on the criterion performances. The former can be readily assessed; it is the latter that is the greater problem for comparing criterion measures.

If two criterion measures are both good measures of the desired criterion performance *and* if individuals have been highly selected directly (via retention and promotion policies) or indirectly (via a valid predictor) for their performance on the criterion, then those two criterion measures will have a low observed correlation. If two criterion measures tap somewhat different dimensions of job performance, they may be differentially restricted in range because the workers may have been selected more strongly for one type of performance than for another. If criterion measures are differentially restricted in range, then the rank order of their observed correlations may not be the same as the rank order of their true correlations in the relevant population, thus providing misleading estimates of which measures are most equivalent in factor structure. For example, a work sample test, a task rating scale, and a job knowledge test may all have equal true correlations with each other in unrestricted samples, but if the first two measures tap a performance dimension that the third does not (say, performance on psychomotor tasks), and if the organization happens to select most strongly for high psychomotor performance, then the observed correlation between the first two measures will be disproportionately low. The more restricted in range a sample is on criterion performances, especially if there is differential restriction in range for alternative measures, the more distorted one's interpretations of the content and relevance of those measures is likely to be.

A major problem with restriction in range on criterion performances is that we typically do not know what the population variance is on any criterion measure and so have no direct basis for correcting for restriction in range. Nor can we collect such data typically, because job performance criterion measures assume that any sample being tested has already been trained, which an applicant or recruit population will not have been.

It is not known to what extent, if any, restriction in range on criterion performances will typically interfere with making appropriate inferences about factorial equivalence.

Criterion Measures Not All Available in Same Sample

Researchers may sometimes want to compare criterion measures that have been used in different studies. For example, an organization may wish

to compare the validation data for one criterion measure to those for a different criterion measure developed elsewhere within or outside the organization. Such comparisons may reflect an effort to synthesize past research by different investigators or an effort to get maximum mileage from limited resources for criterion development.

Making such cross-sample comparisons of different measures is difficult, however, because only predictor-dependent methods of assessing factorial equivalence among all the criterion measures will be available. (Task overlap may be available, but it provides no information about equivalence based on actual criterion performances.) Correlations among all the criterion measures cannot be calculated, which means in turn that no factor analyses of the total criterion space can be conducted. One is required to make indirect comparisons, and these comparisons—say, in item statistics—are further complicated by possible differences across samples in restriction in range on any given dimension of job performance and by the need to determine whether the jobs in question are sufficiently similar in performance demands to be considered the same job or members of the same job family. Assessments of equivalence thus must rely more heavily in these situations on judgments about the nature of the jobs studied and how people have been selected into or out of them.

Predictor-dependent comparison strategies will probably still be available if the studies of the different criterion measures share some common predictors in such cross-study comparisons. If the predictor factor space substantially overlaps all of the criterion measures (as would be indicated by high communalities or high multiple correlations for each criterion measure), then estimates of degree and nature of criterion equivalence probably will be good. Some inference can often be drawn about criterion equivalencies and nonequivalencies when there is less overlap of the predictor factor space with the criterion measures, but it will be difficult to draw any conclusions when the overlap is small. As the overlap with predictors decreases, degree and nature of criterion overlap is less discernible.

It may not always be possible to make strong inferences about the nature and degree of criterion equivalence when criterion measures are examined in separate studies. However, such studies can provide good hypotheses for a second round of validation studies in which criterion equivalence can be directly assessed by collecting all necessary data from the same samples. A second round of validation research could consist of setting up specific and direct tests of those hypotheses using the full complement of criterion measures judged to be useful. Knowledge gained in the earlier research might also be used to improve the old measures or to fashion composites from pieces of the old.

SUMMARY

The criterion problem in performance measurement has evolved from one of developing more adequate measures of job performance to one of developing procedures for comparing the relative utility of alternative measures for a given purpose. This new aspect was referred to here as the problem of assessing the equivalence of criterion measures, where equivalence refers to types and degrees of similarities and differences among criterion measures. Careful evaluation is necessary for developing and selecting the most useful criterion measures; neither psychometric equivalence nor overall utility should ever be assumed.

Five facets of criterion equivalence should be weighed in making a decision to adopt some criterion measures rather than others, or to substitute one for another: relative validity, reliability, susceptibility to compromise, financial cost, and acceptability to interested parties. Although all five facets of equivalence are important, validity is preeminent. Therefore, most of this paper has been devoted to the nature and determination of criterion validity.

Two components of overall criterion validity were described in detail: (1) the construct validity of the criterion measure and (2) the relevance of the performance construct actually measured. Construct validity refers to inferences about the meaning or proper interpretation of scores on a measure and thus requires a determination of just what performance factors are and are not being tapped by a given criterion measure. Relevance refers to the value of differences in criterion performance for promoting the organization's stated goals. It is essential to establish the relevance of criterion measures before deciding which ones to adopt, but relevance seldom can be assessed without first establishing the construct validity (appropriate interpretation) of the criterion performances being measured.

The test development process involves developing a priori hypotheses about the validity, for particular purposes, of the measure under development; validation is a process of empirically testing those hypotheses. Logic, theory, and research all play an important role in these processes, and the higher the quality and quantity of each, the better supported one's inferences about construct validity and relevance will be. Both test development and validation are improved by explicit and detailed accounts of all aspects of the development and validation efforts, from a clarification of the organization's goals for criterion measurement to a description of the data and theory on which judgments about the relevance of a performance construct are based.

The following outline summarizes the process of assessing criterion equivalence that is described in this paper. This outline is presented as only one strategy for analyzing criterion equivalence. Determining criterion equivalence, like

determining the validity of any single criterion, is not a matter of performing some specified procedure. Rather, it is a process of hypothesis testing limited only by the clarity of the organization's goals and by the resources and ingenuity of the investigator.

Outline of a Strategy for Assessing Criterion Equivalence

A. Explicitly specify definitions, hypotheses, and measurement procedures.
   Define organizational goals.
   Define the a priori performance construct.
   State hypotheses about how the performance construct is relevant to the organizational goals.
   Describe procedures used to operationalize the performance construct.
   Describe sample(s) of workers used in the validation research.

B. Do preliminary empirical analyses of properties of individual criterion and predictor measures.
   Estimate reliabilities.
   Estimate degree of restriction in range (empirical estimates possible only for the predictors).
   Compare internal psychometric properties.
   Transform scores where appropriate to equate scaling procedures.

C. State tentative hypotheses about appropriate interpretations (construct validity) of the different criterion measures, based on A and B above.

D. Empirically assess nonequivalencies in construct validity of criterion measures (with disattenuated correlations).

1. Are the criterion scores from the two measures available from the same sample?
   If yes, go to 2 below. If no, go to 3.
2. Is the correlation between two criterion measures > .9?
   If yes, the measures are equivalent in construct validity. Go to 5a.
   If no, go to 5a.
3. Is there differential restriction in range in the predictors?
   If yes, correct for differences in restriction in range. Go to 4.
   If no, go to 4.
4. What are the R^2s when criterion measures are regressed on common predictors (i.e., is it possible to demonstrate equivalence across samples, even when it exists)?
   If both R^2s > .9, equivalence can be determined. Go to 5c. If R^2s are very different, measures are not equivalent. Go to 5c.
   If R^2s are similar but not high, it may not be possible to determine whether equivalent or not. Go to 5c.
5. What is the substantive interpretation of scores on each criterion measure?

a. If criterion measures are numerous, factor analyze the criterion

measures to determine nature of their overlap and nonoverlap in the criterion space. Go to 5b.

b. Relate the criterion factors from 5a above to the predictors (e.g., factor analyze criterion and predictors together, or correlate criterion factors with predictor factors or individual predictors). Go to 5d.

c. Factor analyze the common predictor (if sufficient in number) across different samples with criterion measures added by extension. Go to 5d.

d. Compare patterns of correlations of criterion measures with all available variables. Go to 6.

6. In view of existing measurement limitations, just how strong is the new empirical evidence (from B and D above) relative to the evidence and argument supporting the a priori hypotheses (A above)?

If strong, go to F. If weak, go to E.

E. Perform additional research with existing measures (e.g., with new or larger samples, more predictors, or experimental treatments). Return to A-D, as necessary.

F. State post hoc hypotheses about the appropriate interpretations (construct validity) of the different criterion measures based on B and D above.

G. Reassess the relevance of each criterion measure, based on the revised interpretations in F above.

1. Does it appear possible to improve the relevance of one or more criterion measures (for the organization's particular goals) by improving or combining the measures to better approximate the desired performance construct (which may no longer be the same as in A above).

If yes, return to A. If no, go to H.

H. Compare the overall utility of each criterion measure, weighing their relative: validity (specifically, relevance); reliability; susceptibility to compromise; financial cost; and acceptability to interested parties.

I. Decide about which criterion measure(s), if any, to adopt or substitute for each other.

J. Continue monitoring organizational goals and relevant research, and provide some evaluation of the actual consequences of the decision in H above—all to monitor whether the decision in H should be revised at some point, criterion measures modified, more research done, and so on.

Note: The foregoing strategy provides evidence for meeting many of the applicable American Psychological Association test standards (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985), particularly in Sections 1-3 and 10.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1985 *Standards for Educational and Psychological Testing*. Washington, D.C.: American Psychological Association.

Armor, D.J., R.L. Fernandez, K. Bers, and D. Schwarzbach 1982 Recruit Aptitudes and Army Job Performance: Setting Enlistment Standards for Infantrymen. R-2874-MRAL. Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics), U.S. Department of Defense, Washington, D.C.

Arvey, R.D. 1979 *Fairness in Selecting Employees*. Reading, Mass.: Addison-Wesley.

Astin, A.W. 1964 Criterion-centered research. *Educational and Psychological Measurement* 24 (4):807-822.

Bartlett, C.J. 1983 Would you know a properly motivated performance appraisal if you saw one? Pp. 190-194 in F. Landy, S. Zedeck, and J. Cleveland, eds., *Performance Measurement and Theory*. Hillsdale, N.J.: Erlbaum.

Brogden, H.E., and E.K. Taylor 1950 The theory and classification of criterion bias. *Educational and Psychological Measurement* 10:169-187.

Campbell, J.P. 1983 Some possible implications of "modeling" for the conceptualization of measurement. Pp. 277-298 in F. Landy, S. Zedeck, and J. Cleveland, eds., *Performance Measurement and Theory*. Hillsdale, N.J.: Erlbaum.

Cascio, W.F., and N.F. Phillips 1979 Performance testing: a rose among thorns? *Personnel Psychology* 32:751-766.

Christal, R.E. 1974 The United States Air Force Occupational Research Project. NTIS No. AD774 574. Air Force Human Resources Laboratory (AFSC), Lackland Air Force Base, Tex.

Cronbach, L.J. 1971 Test validation. Pp. 443-507 in R. L. Thorndike, ed., *Educational Measurement*. Washington, D.C.: American Council on Education.

1979 The Armed Services Vocational Aptitude Battery—a test battery in transition. *Personnel and Guidance Journal* 57:232-237.

Cronbach, L.J., G.C. Gleser, H. Nanda, and N. Rajaratnam 1972 *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Curran, C.R. 1983 Comments on Vineberg and Joyner. Pp. 251-256 in F. Landy, S. Zedeck, and J. Cleveland, eds., *Performance Measurement and Theory*. Hillsdale, N.J. : Erlbaum.

Dunnette, M.D. 1976 Aptitudes, abilities, and skills. Pp. 473-520 in M.D. Dunnette, ed., *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally College Publishing Company.

Fleishman, E.A. 1975 Toward a taxonomy of human performance. *American Psychologist* 30:1127-1149.

Fleishman, E.A., and M.K. Quaintance 1984 *Taxonomies of Human Performance: The Description of Human Tasks*. Orlando, Fla.: Academic Press.

Ghiselli, E.E. 1966 *The Validity of Occupational Aptitude Tests*. New York: Wiley.

Gordon, R.A. 1987 Jensen's contributions concerning test bias: a contextual view. In S. Modgil and C. Modgil, eds., *Arthur Jensen: Consensus and Controversy* . Sussex, England: Falmer Press.

Gorsuch, R.L. 1974 *Factor Analysis*. Philadelphia: Saunders.

Gottfredson, L.S. 1984 The Role of Intelligence and Education in the Division of Labor. Report No. 355. Center for Social Organization of Schools, The Johns Hopkins University, Baltimore, Md.

1985 Education as a valid but fallible signal of worker quality: reorienting an old debate about the functional basis of the occupational hierarchy. Pp. 123-169 in A.C. Kerckhoff, ed., *Research in Sociology of Education and Socialization*, Vol.5. Greenwich, Conn.: JAI Press.

1986 Societal consequences of the g factor in employment. *Journal of Vocational Behavior* 29:379-410.

Guion, R.M. 1961 Criterion measurement and personnel judgments. *Personnel Psychology* 14:141-149.

1976 Recruiting, selection, and job placement. Pp. 777-828 in M. D. Dunnette, ed., *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally College Publishing Company.

1978 Scoring of content domain samples: the problem of fairness. *Journal of Applied Psychology* 63:499-506.

1983 The ambiguity of validity: the growth of my discontent. Presidential address to the Division of Evaluation and Measurement at the annual meeting of the American Psychological Association, Anaheim, Calif., August.

1985 Personal communication. October 9.

Gulliksen, H. 1968 Methods for determining equivalence of measures. *Psychological Bulletin* 70:534-544.

Hale, M. 1983 History of employment testing. Pp. 3-38 in A.K. Wigdor and W.R. Garner, eds., *Ability Testing: Uses*, *Consequences*, *and Controversies*. *Part II: Documentation Section*. Washington, D.C.: National Academy Press.

Hunter, J.E. 1983 A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. Pp. 257-266 in F. Landy, S. Zedeck, and J. Cleveland, eds., *Performance Measurement and Theory*. Hillsdale, N.J.: Erlbaum.

Hunter, J.E., and F.L. Schmidt 1983 Quantifying the effects of psychological interventions on employee job performance and work-force productivity. *American Psychologist* 38:473-478.

Hunter, J.E., F.L. Schmidt, and J. Rauschenberger 1984 Methodological, statistical, and ethical issues in the study of bias in psychological tests. Pp. 41-99 in C.R. Reynolds and R.T. Brown, eds., *Perspectives on Bias in Mental Testing*. New York: Plenum.

Ironson, G.H., R.M. Guion, and M. Ostrander 1982 Adverse impact from a psychometric perspective. *Journal of Applied Psychology* 67:419-432.

James, L.R. 1973 Criterion models and construct validity for criteria. *Psychological Bulletin* 80 (1):75-83.

Jenkins, J.G. 1946 Validity for what? *Journal of Consulting Psychology* 10:93-98.

Jensen, A.R. 1980 *Bias in Mental Testing*. New York: Free Press.

1985 Armed Services Vocational Aptitude Battery (test review). *Measurement and Evaluation in Counseling and Development* 18:32-37.

1987 The g beyond factor analysis. In J.C. Conoley, J.A. Glover, and R.R. Renning, eds., *The Influence of Cognitive Psychology on Testing and Measurement*. Hillsdale, N.J.: Erlbaum.

Landy, F.J. 1986 Stamp Collecting vs. Science: Validation as Hypothesis Testing. *American Psychologist* 41:1183-1192.

Landy, F.J., and J.L. Farr 1983 *The Measurement of Work Performance: Methods*, *Theory*, *and Applications* . New York: Academic Press.

Landy, F., S. Zedeck, and J. Cleveland, eds. 1983 *Performance Measurement and Theory*. Hillsdale, N.J.: Erlbaum.

Messick, S. 1975 The standard problem: meaning and values in measurement and evaluation. *American Psychologist* 30:955-966.

Muckler, F.A. 1982 Evaluating productivity. Pp. 13-47 in M.D. Dunnette and E.A. Fleishman, eds., *Human Performance and Productivity: Human Capability Assessment* . Hillsdale, N.J.: Erlbaum.

Nagle, B.F. 1953 Criterion development. *Personnel Psychology* 6:271-289. Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics)

1983 Second Annual Report to the Congress on Joint-Service Efforts to Link Standards for Enlistment to On-the-Job Performance. A report to the House Committee on Appropriations, U.S. Department of Defense, Washington, D.C.

Osborn, W. 1983 Issues and strategies in measuring performance in army jobs. Paper presented at the annual meeting of the American Psychological Association, Anaheim, Calif.

Pickering, E.J., and A.V. Anderson 1976 Measurement of Job-Performance Capabilities. TR 77-6. Navy Personnel Research and Development Center, San Diego, Calif.

Ree, M.J., C.J. Mullins, J.J. Mathews, and R.H. Massey 1982 Armed Services Vocational Aptitude Battery: Item and Factor Analyses of Forms 8, 9, and 10. Air Force Human Resources Laboratory (Manpower and Personnel Division, AFSC), Lackland Air Force Base, Tex.

Richards, J.M., Jr., C.W. Taylor, P.B. Price, and T.L. Jacobsen 1965 An investigation of the criterion problem for one group of medical specialists. *Journal of Applied Psychology* 49:79-90.

Schmidt, F.L. 1977 The Measurement of Job Performance. U.S. Office of Personnel Management, Washington, D.C.

Schmidt, F.L., and J.E. Hunter 1981 Employment testing: old theories and new research findings. *American Psychologist* 36:1128-1137.

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

Schmidt, F.L., J.E. Hunter, and V.W. Urry 1976 Statistical power in criterion-related validity studies. *Journal of Applied Psychology* 61:473-485.

Schmidt, F.L., J.E. Hunter, and K. Pearlman 1981 Task differences as moderators of aptitude test validity in selection: a red herring. *Journal of Applied Psychology* 66:166-185.

Schmidt, F.L., J.E. Hunter, and A.N. Outerbridge 1985 The Impact of Job Experience and Ability on Job Knowledge, Work Sample Performance, and Supervisory Ratings of Job Performance . U.S. Office of Personnel Management, Washington, D.C.

Schmidt, F.L., A.L. Greenthal, J.E. Hunter, J.G. Berner, and F.W. Seaton 1977 Job sample vs. paper-and-pencil trades and technical tests: adverse impact and examinee attitudes. *Personnel Psychology* 30:187-197.

Schoenfeldt, L.F. 1982 Intra-individual variation and human performance. Pp. 107-134 in M.D. Dunnette and E.A. Fleishman, eds., *Human Performance and Productivity: Human Capability Assessment*. Hillsdale, N.J.: Erlbaum.

Severin, D. 1952 The predictability of various kinds of criteria. *Personnel Psychology* 5:93-104.

Sinden, J.A., and A.C. Worrell 1979 *Unpriced Values: Decisions Without Market Prices*. New York: Wiley.

Smith, P.C. 1976 Behaviors, results, and organizational effectiveness: the problem of criteria. Pp. 745-775 in M.D. Dunnette, ed., *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally College Publishing Company.

1985 Global measures: do we need them? Address presented at the annual meeting of the American Psychological Association, Los Angeles, August.

Smith, P.C., L.M. Kendall, and C.L. Hulin 1969 *The Measurement of Satisfaction in Work and Retirement*. Chicago: Rand McNally.

Staw, B.M. 1983 Proximal and distal measures of individual impact: some comments on Hall's performance evaluation paper. Pp. 31-38 in F. Landy, S. Zedeck, and J. Cleveland, eds., *Performance Measurement and Theory* . Hillsdale, N.J.: Erlbaum.

Staw, B.M., and G.R. Oldham 1978 Reconsidering our dependent variables: a critique and empirical study. *Academy of Management Journal* 21:539-559.

Tenopyr, M.L. 1977 Content-construct confusion. *Personnel Psychology* 30:47-54.

1985 Test and testify: can we put an end to it? Address presented at the annual meeting of the American Psychological Association, Los Angeles, August.

Uhlaner, J.E., and A.J. Drucker 1980 Military research on performance criteria: a change of emphasis. *Human Factors* 22:131-139.

U.S. Department of Defense 1984 Test Manual for the Armed Services Vocational Aptitude Battery. United States Military Entrance Processing Command, 2500 Green Bay Road, North Chicago, Ill. 60064.

U.S. Department of Labor 1970 Manual for the USTES General Aptitude Test Battery. Manpower Administration, U.S. Department of Labor, Washington, D.C.

Vineberg, R., and J.N. Joyner 1983 Performance measurement in the military services. Pp. 233-250 in F. Landy, S. Zedeck, and J. Cleveland, eds., *Performance Measurement and Theory*. Hillsdale, N.J.: Erlbaum.

Wallace, S.R. 1965 Criteria for what? *American Psychologist* 20:411-417.

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

Wherry, R.J., and C.J. Bartlett 1982 The control of bias in ratings: a theory of rating. P*ersonnel Psychology* 35:521551.

Wherry, R.J., P.F. Ross, and L. Wolins 1956 A Theoretical and Empirical Investigation of the Relationships Among Measures of Criterion Equivalence. NTIS No. AD 727273. Research Foundation, Ohio State University, Columbus.

Wigdor, A.K., and W.R. Garner, eds. 1982 *Ability Testing: Uses*, *Consequences*, *and Controversies*. *Part I: Report of the Committee*. Washington, D.C.: National Academy Press.

Range Restriction Adjustments in the Prediction of Military Job Performance

Stephen B. Dunbar and Robert L. Linn

INTRODUCTION

Common practice in establishing the criterion-related validity of a test or battery of tests to be used for selection or classification involves performing a linear regression of a relevant measure of performance on the test battery and reporting various descriptive statistics that assess the magnitude of the linear relationship between predictor(s) and the criterion. Although alternatives to the familiar correlation coefficient and perhaps less familiar regression slopes and intercept exist, these alternatives make the task of characterizing criterion-related validity of the battery more cumbersome. The correlation coefficient, in particular, allows for ready comparisons of predictive validity across occupational categories as well as across different predictor and criterion measures, so that its widespread use is not surprising. In the context of military performance assessment, correlations allow for comparisons of predictive validity over time and across Services. Such comparisons are important components in the validation of selection composites that are used for prediction in a wide variety of occupational categories such as military training programs.

Whatever appeal correlations have for these and other reasons must be weighed against certain limitations, several of which are especially relevant in the context of criterion-related validity studies. When a correlation coefficient is used to make inferences about the predictive validity of a test

battery for a population of applicants or enlistees, its values should ideally be estimated from a random sample of the applicant population. In most settings where predictive validity needs to be established, the only sample available for estimating the desired population values is one containing individuals, selected in part on the basis of scores on the predictor(s) in question, who have remained with a program long enough for criterion scores to be obtained. Besides the difficulties presented by the use of nonrandom samples in calculating correlations, the well-known sensitivity of correlations, slopes, and intercepts to linearity and homoscedasticity in the joint distribution of predictors and criteria is an area of concern. This concern can be magnified by selection effects.

This paper provides an overview of standard procedures used to adjust correlations and regression parameters for the effects of selection, commonly referred to as corrections for range restriction. Technical issues related to the accuracy of these adjustments are considered, especially where they are likely to have implications for the types of adjustment procedures appropriate for large-scale predictive validity studies of an aptitude battery like the Armed Services Vocational Aptitude Battery (ASVAB). The paper concludes with a discussion of issues related to the implementation of a set of adjustment procedures for validation studies in the military, where the choice of the reference population, choice of selection variables for making adjustments, and choice of an analytical procedure all have important consequences for the assessment of the predictive validity of present and future versions of the ASVAB.

SELECTION EFFECTS IN CORRELATION AND REGRESSION

Although the effects of various types of nonrandom selection on correlation coefficients, slopes, and intercepts are well-documented in the psychometric literature (cf. Thorndike, 1949; Gulliksen, 1950; Lord and Novick, 1968), a brief review of these effects will establish the context for technical issues related to their use in studies of criterion-related validity. Figures 1-4 illustrate the effects of the usual types of selection on the bivariate scatterplot of a selection test (X) and a performance criterion (Y). In Figure 1, the scatterplot of a sample of 5,000 observations from a bivariate normal population is shown, along with the least-squares regression line of Y on X. The correlation between X and Y in this population is .60. Figures 2 and 3 demonstrate the effects of *explicit* selection on X and Y, respectively—explicit selection in these examples involves actual truncation of the marginal distribution of the selection variable and is clearly visible by inspection of the scatterplot. When selection is explicit on X, the Y on X least-squares regression line is unaltered because of assumed linearity. However, estimates of the correlation between X and Y are altered because of the reduced

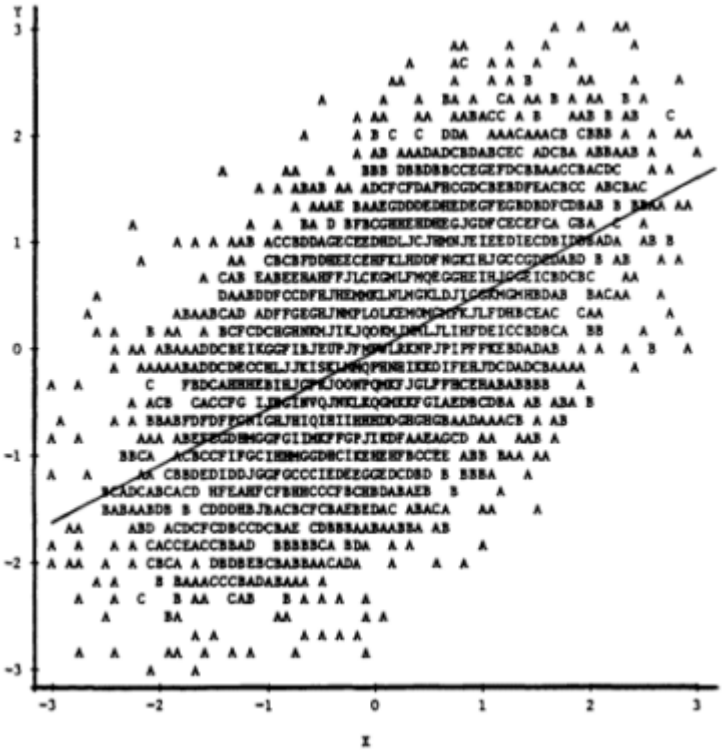


Figure 1 Scatterplot of a criterion (*Y*) and a predictor (*X*) in an unselected sample (N = 5,000, ρ_{xy} = .60).

variance of *X* in the selected sample. In Figure 2, with only the upper 50 percent of the observations on *X* included, the resulting correlation of .398 underestimates the population correlation by a substantial amount. When selection is explicit on *Y*, both the terms in the *Y* on *X* least-squares regression and the correlation coefficient are affected. In Figure 3, with only the upper 50 percent of scores on *Y* included, the regression line (depicted by the broken line) has a smaller slope and larger intercept, while the correlation between *X* and *Y* of .412 again underestimates the population value.

While explicit selection can occur, particularly when *X* is a screening device like an admissions test, a more likely situation would depict *X* as one of several measures used to select individuals. In such a situation one might imagine a third variable, *Z*, as the explicit selection variable—*Z* can be

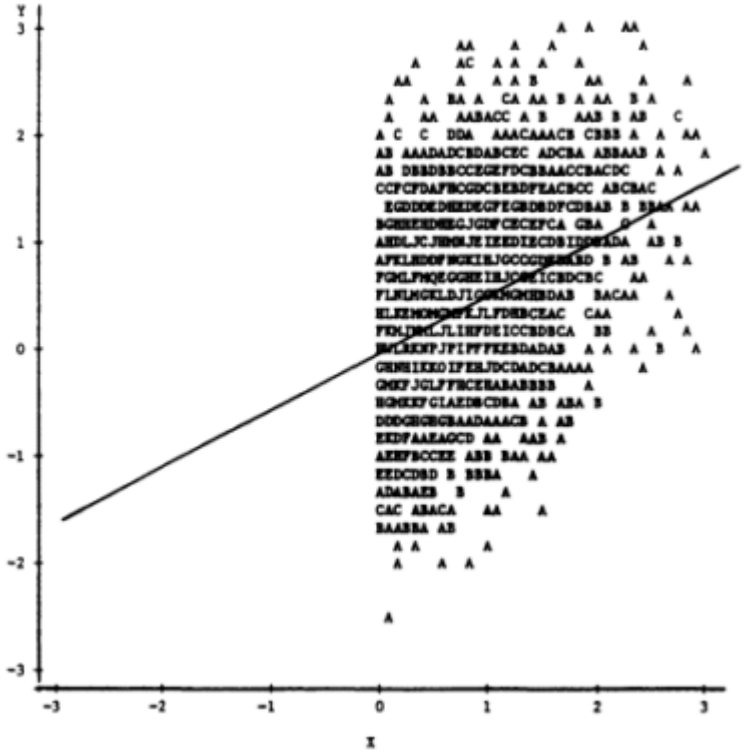


Figure 2 Explicit selection on *X*.

thought of as a kind of composite measure that includes such factors as self or administrative selection as well as scores on other predictors and is positively correlated with *X* and *Y* under typical circumstances. That *Z* can be thought of as a composite measure is reflected in its designation as an ideal discriminant function separating selected and nonselected groups (Cronbach et al., 1977) or as a latent variable underlying the true selection process (Muthén and Jöreskog, 1983). In this case, both *X* and *Y* are referred to as *incidental* selection variables because the selection effects on the bivariate distribution of *X* and *Y* are indirect. This type of selection effect also exists when interest focuses on the predictive value of an alternative set of variables imperfectly correlated with the explicit selection variable. In such a situation, the alternative predictors represent incidental selection variables.

The more subtle effects of explicit selection on *Z* are illustrated in Figure 4. The *XY* scatterplot in this figure is based on a trivariate normal population

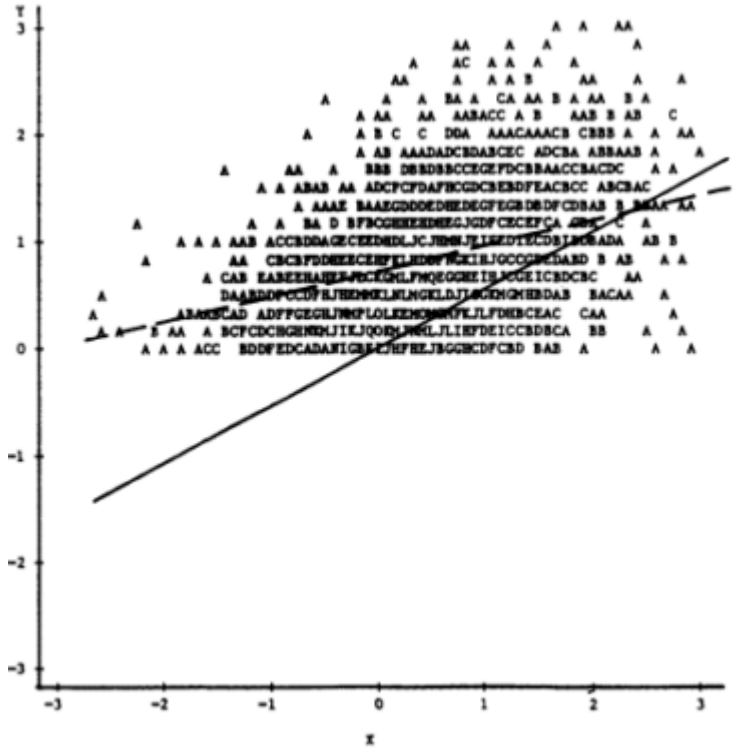


Figure 3 Explicit selection on *Y*.

distribution of *X*, *Y*, and *Z* in which the correlation between *Y* and each of *X* and *Z* is .60 and the correlation between *X* and *Z* is .90. The scatterplot contains only the upper 50 percent of the observations on *Z*, and, as can be seen in the figure, the only evidence for this being a selected sample is a reduction in the variability of the marginal distributions of *X* and *Y*. Even when *X* and the true selection variable are as highly correlated as in this example, the effects of incidental selection yield a least-squares regression line (again depicted by the broken line) with reduced slope and a correlation between *X* and *Y* (.414) that underestimates the population correlation substantially.

Although the plots in Figures 1-4 do provide an indication of the types of selection effects that can occur in practice, they do not show how differential selectivity can complicate the interpretation of correlations based on selected samples. Table 1 illustrates the effects of different degrees of range

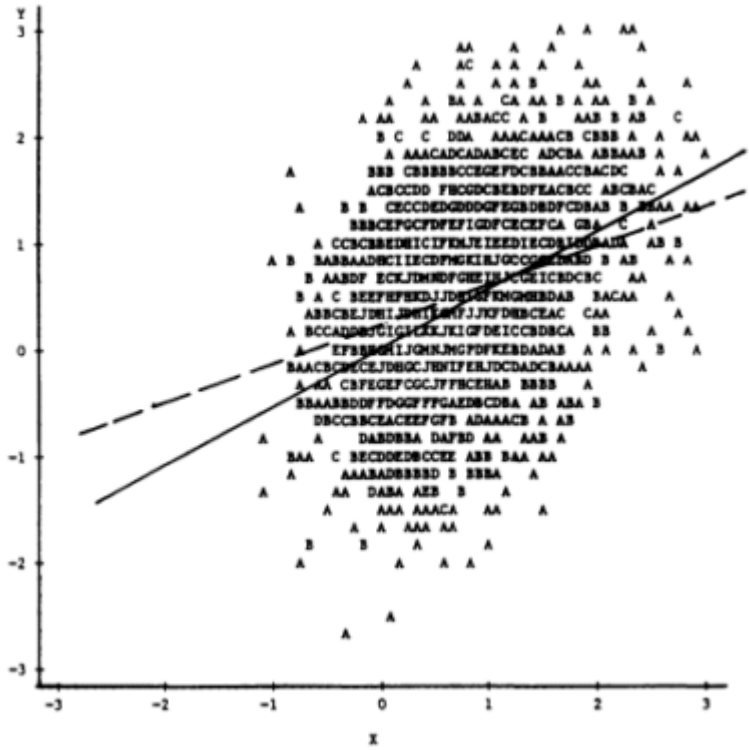


Figure 4 Explicit selection on *Z*.

restriction on the correlations of two measures with a criterion; one of the two measures represents the explicit selection variable (*Z* in the above discussion), while the other represents an incidental selection variable (*X*). The statistics in the table are based on a trivariate normal population in which the correlations between the criterion, *Y*, and i and *X* are .50 and .45, respectively, and the correlation between *Z* and *X* is .50. The selection ratio indicates the proportion of the total sample that is selected into the validation sample on the basis of scores on the explicit selection variable *Z*. The standard deviation of *Z* is given in the second column of the table and the correlations between the predictor variables and the criterion in the third and fourth columns.

With no selection on *Z*, the correlations shown in the table are equal to the population values of .50 and .45. However, as the selection ratio decreases

TABLE 1 Illustration of the Effects of Range Restriction for Z on the Correlation of Z and X with a Criterion Measure Y

Selection	Standard	Correlation with *Y*	
Ration	Deviation of *Z*	Z	X
1.0	1.000	.50	.45
.7	.701	.38	.37
.6	.649	.35	.36
.5	.603	.33	.35
.4	.559	.31	.34
.3	.515	.29	.33
.2	.468	.26	.32
.1	.411	.23	.31
.05	.371	.21	.30

NOTE: Calculations are based on an assumed trivariate normal distribution with a correlation of .50 between *Z* and *X* in the unselected population.

and the proportion of individuals excluded from the validation sample increases, the correlations between the predictors and the criterion steadily decrease. For example, if only individuals in the top 70 percent of scores on *Z* are selected (selection ratio of .7) and hence available for a validation study, the estimates of the correlations of *Z* and *X* with the criterion would be expected to decrease to .38 and .37, respectively. When only the upper 30 percent are selected, the corresponding correlations become .29 and .33, while for selection of the upper 5 percent they become .21 and .30. The effects of explicit and incidental selection shown in Table 1 clearly lead to a steady decrease in the assessed predictive values of *Z* and *X* when product-moment correlations are used to describe the degree of association with the criterion. However, the table also shows this decrease to be more dramatic for the explicit selection variable, *Z*, than it is for *X*. In spite of the fact that the population correlations indicate *Z* to be the better predictor, the effects of explicit selection on *Z* lead to a misleading indication that *X* is the better predictor once the selection ratio drops below .6. The degree by which one is misled clearly increases as the degree of selectivity increases. Examples such as these provide a clear indication that understanding of selection effects is crucial to the interpretation of results from criterion-related validity studies. Without risk of hyperbole, one could say that the predictive validity of a selection instrument cannot be accurately characterized unless the possible effects of sample selection are accounted for to some extent.

Coping With the Effects of Selection

The subtle ways in which biases introduced by sample selection affect estimates of correlations in an entire applicant pool are difficult to account for in any exhaustive way; however, it is possible to obtain useful assessments of the degree of association between a set of predictors and a criterion even with selected samples. Although no method of adjusting for selection effects is without limitations, it is often possible to obtain a less biased indication of predictive validity either by employing an adjustment procedure or by examining alternatives to the validity coefficient.

Allred (in this volume) provides an excellent review of alternatives to the validity coefficient for describing the results of a predictive validity study, many of which are less sensitive to the effects of selection. The scatterplots shown previously are the most straightforward example of an alternative approach. They provide a very detailed representation of the relationship between a predictor and a criterion measure. Such detail can be important for detecting specific characteristics of the relationship-ceiling and floor effects, marked departures from linearity, changes in the degree of criterion variability depending on predictor score (heteroscedasticity), and outliers can all be discerned from careful inspection of a scatterplot. In addition to the scatterplot, Allred shows how graphical displays of the criterion distribution at fixed levels of the predictor can offer a more concise evaluation of potential nonlinearity and heteroscedasticity than the scatterplot. The various types of expectancy tables discussed by Allred provide estimates of the criterion performance expected for any prespecified level of performance on the predictor(s) and are also useful indicators of the predictor-criterion relationship. General measures of association used in the analysis of contingency tables (cf. Bishop et al., 1975) could conceivably be applied to such expectancy tables in order to provide a numerical index analogous to the correlation coefficient—but the tables are also valuable in their own right.

When scatterplots or plots of conditional distributions indicate that the relationship between predictor and criterion is approximately linear and homoscedastic, the use of the familiar summary statistics from correlation and regression is most meaningful. Although a bivariate, or in the case of multiple predictors multivariate, normal distribution of predictor(s) and criterion is not strictly required for these conditions to hold, normality is the basis for the significance tests and confidence intervals used for inferences about population correlations and regression parameters. The obvious difficulty for formal statistical inference in the context of a criterion-related validity study lies in the fact that the random sampling scheme required, given that normality conditions are satisfied, can seldom be attained in practice. The estimates of the unknown population values are biased by nonrandom sample selection—the adjustment procedures reviewed below

attempt to remove at least a portion of this bias by incorporating specific assumptions about the selection process into estimates of correlations and regression parameters.

Corrections for Sample Selection Bias

The most common procedures for adjusting correlations for the effects of sample selection were first introduced by Pearson (1903) for the bivariate case and later extended by Lawley (1943-1944) to the case of multiple predictors and criteria. These procedures are sometimes referred to as the Pearson-Lawley corrections and have been used extensively in validation research in certain Services for a number of years. As discussed by Lord and Novick (1968), Lawley's extension of Pearson's two-and three-variable corrections describes the relationships between complete sets of explicit and incidental selection variables based on two assumptions: (1) the regression of each incidental selection variable on any combination of the explicit selection variables is linear, and (2) the errors of estimate incurred in regressing incidental on explicit selection variables are constant (i.e., they are homoscedastic). When these conditions are satisfied, the covariances among the incidental selection variables can be expressed as

$$C_{xx} = C^*_{xx} - C^*_{xz}(C^{*-1}_{zz} - C^{*-1}_{zz}C_{zz}C^{*-1}_{zz})C^*_{zx}$$

where C represents the variance-covariance matrix of the variables indicated by subscripts x and z, which refer to incidental and explicit selection variables, respectively. An asterisk is used to distinguish matrices based on the selected sample from those based on the unselected population.

As can be seen from the above expression, it is the relationship among explicit selection variables in the unselected population versus selected sample (C_{zz} vs. C^*_{zz}) that determines the size of the adjustment made by the Lawley correction procedure. For the case of no selection, C_{zz} and C^*_{zz} are identical and the term in parentheses vanishes, yielding $C_{xx} = C^*_{zz}$. As selection affects the elements of C^*_{zz}, the elements of C^*_{xx} are adjusted accordingly.

For the special case illustrated earlier, in which Z was the lone explicit selection variable and X and Y were incidental selection variables of substantive interest, the Pearson-Lawley expression for the correlation between X and Y in the unselected population is given by

$$R_{xy} = \frac{r_{xy} + (S_z^2 / s_z^2 - 1)r_{xz}r_{yz}}{\left\{\left[1 + (S_z^2 / s_z^2 - 1)r_{xz}^2\right]\left[1 + (S_z^2 / s_z^2 - 1)r_{yz}^2\right]\right\}^{1/2}}$$

where upper-case *R* and *S* designate correlations and standard deviations in the unselected population, respectively, and lower-case *r* and *s* designate corresponding quantities in the selected sample. If the explicit selection variable, *Z*, were known, its standard deviation in the unselected population or applicant pool is all that would be necessary in addition to selected sample statistics in order to estimate the *XY* correlation. Thus, the expression given above would be applicable in a situation where, for example, selection was explicit on a composite variable, but interest in the correlations between individual elements of the composite and the criterion existed. Where a number of explicit selection variables are known, the Lawley extension of the above three-variable formula would provide the appropriate adjustment under assumptions (1) and (2).

Inspection of the above formula provides some insight into the concepts underlying this and other corrections for range restriction. The numerator of this formula shows that the sample r_{xy} is being incremented by a factor related to the selection ratio and the magnitudes of the correlations between the explicit selection variable and the variables of substantive interest. As selectivity increases, the ratio of standard deviations in the unselected to selected groups becomes larger and the correction factor increases. Similarly, when the explicit selection variable is highly correlated with the variables of substantive interest, the correction factor can be quite large. In either case, the similarity of the above formula to the one used for calculating partial correlations makes clear that the adjustment for explicit selection on *Z* "undoes" precisely what partial correlation is designed to do. That is, instead of factoring out variance that is not considered related to the correlation between *X* and *Y*, the above equation factors in variance that *is* considered related to that correlation.

If *X* itself happens to be the explicit selection variable, the Pearson two-variable correction formula yielding R_{xy} can be obtained by simply substituting *x* for *z* in the subscripts of the above formula and simplifying the resulting expression, yielding

$$R_{xy} = \frac{(S_x / s_x) r_{xy}}{\left[1 + (S_x^2 / s_x^2 - 1) r_{xy}^2\right]^{1/2}}$$

where upper-and lower-case quantities are defined as before. Because *Z*, conceived as a true selection variable, is not likely to be observed in practice, the above two-variable correction formula has been widely used, even when selection is not, strictly speaking, explicit on *X*.

The application of these formulas is quite simple and can be illustrated by again considering the scatterplots in Figures 2 and 4. Recall that these plots depicted explicit selection on *X* and *Z*, respectively. When selection is

explicit on X, the two-variable selection formula gives a corrected estimate of .577 for the correlation between X and Y in contrast to the observed r_{xy} of .398 in the selected sample. Similarly, when selection is explicit on Z, the three-variable formula gives a corrected estimate of .590 for R_{xy} in contrast to the r_{xy} of .414 in the selected sample. As expected, the corrected values in these two instances are quite close to the population correlation of .60 in that assumptions (1) and (2) are perfectly satisfied and the correct explicit selection variable is available. Perhaps a more realistic example would depict X as the only available proxy for the unobserved true selection variable and treat it as explicit even though it is incidental. When the two-variable formula is thus used in the setting depicted in Figure 4, the corrected estimate of R_{xy} is .540, still smaller than the population correlation. As suggested below, this undercorrection is likely to be a common occurrence when the Pearson-Lawley procedures are used in practice.

Technical Considerations and the Accuracy of Pearson-Lawley Corrections

The principal technical issues that have implications for the value of corrections for range restriction in practice relate to their accuracy in the presence of violated assumptions and their degree of sampling error. With regard to the former area of concern, studies by Rydberg (1963), Linn (1968), Novick and Thayer (1969), Brewer and Hills (1969), Greener and Osburn (1979, 1980), Linn et al. (1981), Dunbar (1982), Gross and Fleischmann (1983), and Booth-Kewley (1985) have addressed the general issue of bias in results of range restriction adjustments when either regression or selection assumptions are not satisfied. Gullickson and Hopkins (1976), Forsyth (1971), Bobko and Rieck (1980), Dunbar (1982), Gross and Kagen (1983), Brandt et al. (1984), and Allen and Dunbar (1990) have provided descriptions of the sampling behavior of estimates of correlations and regression parameters that have been corrected for range restriction. The principal findings of some of these studies are reviewed below.

The concern regarding the effects of nonlinearity and/or heteroscedasticity in the criterion-predictor relationship is aptly expressed by Lord and Novick (1968), who argue that the Pearson-Lawley corrections can give overly optimistic indications of predictive validity when assumptions are not met. In spite of a common observation that departures from linearity and homoscedasticity are likely to occur at the same time in applied settings, most studies of the behavior of the Pearson-Lawley corrections under violated assumptions have examined nonlinearity and heteroscedasticity separately. Results of studies using both real and simulated data suggest that the effects of violated assumptions on the accuracy of the Pearson-Lawley corrections depend on the nature of the violation. Greener and Osburn

(1979, 1980), for example, found nonlinearity to be a more serious concern than heteroscedasticity for low to moderate degrees of selectivity when using Pearson's two-variable correction for explicit selection. These authors found a strong tendency toward undercorrection when the slope of the regression line was smaller in the extreme portions of the predictor scale than it was in the middle. Moreover, the magnitude of the undercorrection increased as the degree of selectivity increased. Dunbar (1982) obtained a similar result with simulated data, but also found that overcorrections could occur when the slope was higher in the extremes of the predictor scale than in the middle.

Although the effects of nonlinearity can be quite severe, it was a concern for heteroscedasticity in the joint predictor-criterion distribution that led Lord and Novick (1968) to their words of caution. Test scores in particular, they argued, tend to have distributional forms such that variation about a regression line is apt to be smallest in the tails. When only the upper 10 or 20 percent of an applicant group is selected, an inflated estimate of the population correlation can result. Novick and Thayer (1969) documented this effect for the case of explicit selection in an empirical study of the accuracy of the Pearson-Lawley correction formulas. Using real data sets truncated from the left to represent various degrees of range restriction, they found corrections to have uncertain precision for extreme degrees of selection and a preponderance of overcorrections in sets of data where the principal violation was the presence of heteroscedasticity. Greener and Osburn (1980) extended this finding through a more systematic set of simulated bivariate distributions in which the scatter around the regression line was reduced at the extremes of the predictor score distribution. They too found overcorrections to predominate when the selected sample represented less than half of the unselected population.

Booth-Kewley (1985) investigated the accuracy of univariate and multivariate corrections using ASVAB test scores and criterion data from seven Navy technical training schools. Time, expressed as number of days that a student took to complete the training course, served as the criterion variable. Validity coefficients of ASVAB tests for all students completing courses were used as unrestricted "population" values for each course. Restricted samples with selection ratios of .10 through .90 in steps of .10 were created by truncation on the selection composite for each school. Multivariate corrections were generally closer to the unrestricted validities than were the univariate corrections, and the latter were still generally better than the uncorrected values. Overcorrections were common for both the univariate and multivariate procedures, however, perhaps because of the distributions of the criterion variable, time till completion.

Results that suggest the Pearson-Lawley correction formulas might lead to systematic overestimation of the predictor-criterion correlation are especially

disturbing when one considers their practical application in educational or training programs in which the exact character of the joint predictor-criterion distributions cannot be determined. In such settings, however, it is likely that selection is not explicitly made on the basis of an available predictor or set of predictors, but is based on some kind of composite measure that is not perfectly correlated with available selection tests. If the selection tests are truly incidental selection variables, then in evaluating the above tendencies toward overcorrections under particular circumstances, one must consider the dual effect of violated assumptions regarding the selection process as well as the regressions of explicit on incidental selection variables. Linn and colleagues (Linn, 1968; Linn et al., 1981) have illustrated how substituting an incidental selection variable for an explicit selection variable in Pearson's two-variable formula results in a consistent bias toward undercorrection when other assumptions are satisfied and the correlation between the available and true selection variables is at least moderately positive. Whether or not this bias is sufficient to overshadow the potential for overcorrection in the situations noted above was examined by Dunbar (1982) and Gross and Fleischmann (1983). An example from the former illustrates some typical findings with regard to this question.

Dunbar's (1982) examination of the Pearson-Lawley corrections focused on the combined effects of violated assumptions and incompletely specified selection rules (i.e., selection rules where one or more factors involved in sample selection was ignored in the application of a correction formula). While selection was always based on a third variable, it was modeled by both a step function (as in explicit selection on Z) and a smooth probability function (as when an unknown factor like self-selection enters the selection process, making Z itself an incidental selection variable). Table 2 provides a summary of results from applying Pearson's two-and three-variable formulas to selected samples from four types of distributions of X, Y, and Z, such that

(1) All regressions are linear and homoscedastic.
(2) The regression of Y on Z and X has reduced slope at extreme predictor scores.
(3) The regression of Y on Z and X has reduced scatter about the regression line at extreme predictor scores.
(4) The regression of Y on Z and X has both reduced slope and scatter at extreme scores.

Case (2) was intended to provide a situation where undercorrection was likely, case (3) where overcorrection was likely, and case (4) where either under-or overcorrection could occur depending on how the effects of nonlinearity and heteroscedasticity interacted. Entries in the table, except where noted, represent average differences over 80 replications between sample correlations

TABLE 2 Average Differences Between Corrected and Population Correlations in Fisher's Z Equivalents

Selection		Type of *Y* on *X* Regression			
Rule	Ratio	(1)	(2)	(3)	(4)
		Two-Variable Corrections			
Step	.90	-.028	-.029	-.044	-.036
Function	.50	-.118	-.124	-.122	-.130
	.10	-.148	-.190	-.107	-.170
Smooth	.49	-.048	-.048	-.050	-.051
Curve	.37	-.112	-.123	-.113	-.127
	.09	-.142	-.191	-.104	-.171
		Three-Variable Corrections			
Step	.90	-.001	.006	-.006	.000
Function	.50	.001	-.018	.002	-.020
	.10	-.006	-.124	.043	-.093
Smooth	.49	.001	-.002	.001	-.004
Curve	.37	.000	-.026	.004	-.025
	.09	.008	-.127	.042	-.094

NOTE: Averages are over 80 replications except where selection ratios are .10 or .09, in which case 62 and 61 replications were performed, respectively. Type of *Y* on *X* regressions are:

(1) linear and homoscedastic,
(2) nonlinear, with reduced slope at extreme *X* scores,
(3) heteroscedastic, with reduced scatter at extreme X scores, and
(4) nonlinear and heteroscedastic, with reduced slope and scatter at extreme *X* scores.

SOURCE: Based on Dunbar (1982).

corrected for range restriction and population correlations, expressed as Fisher's *Z* equivalents.

Several striking patterns in the behavior of the Pearson-Lawley corrections can be discerned from the results in the table. First of all, when the two-variable correction was used in the absence of information concerning *Z*, undercorrections occurred regardless of the degree to which assumptions were satisfied. The bias toward undercorrection was larger for case (2); however, the dominant pattern is one of increasing negative bias as the selection ratio gets smaller. Of particular note in this regard was that for the heteroscedastic case (3), Dunbar's results indicated that the negative bias introduced by using the two-variable formula outweighed any positive bias introduced by reduced scatter about the regression line.

A second result that is noteworthy involved the interaction between nonlinearity and heteroscedasticity. When these assumptions were violated at the same time, the reduced slope at extreme *X* scores appeared to have a greater

effect on the accuracy of the corrected values than did reduced scatter, making undercorrections common even when the correct three-variable adjustment procedure was used. The dominant influence of nonlinearity was also noted by Gross and Fleischmann (1983), who also provide an important illustration of large overcorrections when the slope of the *Y*-on-*X* regression increased and the scatter about the regression line decreased with increasing predictor scores. However, Dunbar's results suggest that, to the extent reduced slopes at extreme predictor scores are more common in practice, the tendency toward underestimating the validity of a test considered to be only part of a selection decision is stronger than any tendency toward overestimation using the Pearson-Lawley corrections. This is not to say that overestimation will not occur, but rather that it is a bias less likely to occur in typical settings where predictive validity studies are conducted. Because some types of violations can lead to serious overcorrections, standard practice should include investigations of the plausibility of assumptions and the likely nature of any departures from linearity and homoscedasticity. Scatterplots, plots of conditional distributions of the criterion, and plots of residuals, such as those described by Allred (in this volume), can be useful in this regard.

The sampling behavior of correlation coefficients corrected for restriction of range is a second major area of concern in applications of the Pearson-Lawley adjustment procedures. As with any statistical procedure whose use involves some degree of uncertainty, that uncertainty clouds the precision of the resulting statistic as an estimate of a population parameter. As mentioned previously, the sampling error of adjusted correlations has been investigated both empirically (Forsyth, 1971; Gullickson and Hopkins, 1976; Dunbar, 1982; Gross and Kagen, 1983; Brandt et al., 1984 , and Allen and Dunbar, 1990) and analytically (Kelley, 1923; Cohen, 1955; Bobko and Rieck, 1980; Allen and Dunbar, 1990), with results in general agreement that standard errors of corrected correlations can be quite large under conditions of extreme selectivity.

Bobko and Rieck (1980) present an expression discussed by Kelley (1923) for the large-sample standard errors of correlation coefficients corrected for explicit selection. In the case of explicit selection on the predictor, the approximate standard error of the corrected correlation under bivariate normality is a function of the ratio of the standard deviation in the unselected population to that in the selected sample, the magnitude of the correlation in the selected sample, and the sample size. Table 3 illustrates the relative magnitudes of the standard errors of uncorrected, SE(r), and corrected, SE(R), correlations for situations in which the corrected correlation is equal to .5, the size of the sample is 100, and the variables follow a bivariate normal distribution. Given in the columns of the table are values of the ratios of standard deviations, uncorrected and corrected correlations, and their estimated standard errors. In addition, the last two columns of the table indicate,

TABLE 3 Illustrative Values of Standard Errors of Correlations Equal to .5 after Correcting for Explicit Selection on the Predictor

K	r	SE(*r*)	*R*	SE(*R*)	*W*	N'
1.0	.500	.075	.500	.075	1.000	00
1.2	.434	.081	.500	.086	1.065	14
1.4	.381	.085	.500	.098	1.151	35
1.6	.339	.088	.500	.110	1.249	58
1.8	.305	.091	.500	.123	1.354	83
2.0	.277	.092	.500	.135	1.465	117
2.2	.254	.094	.500	.148	1.579	147
2.4	.234	.095	.500	.160	1.696	185
2.6	.217	.095	.500	.173	1.815	232
2.8	.202	.096	.500	.186	1.936	275
3.0	.189	.096	.500	.198	2.058	328

NOTE: A sample of 100 in the selected group is assumed for all SE's. The following notation is used for column headings:
K is the ratio of the standard deviation in the unselected population to the standard deviation in the selected group.
r is the correlation in the selected group.
SE(*r*) is the estimated standard error of r.
R is the correlation corrected for explicit selection on the predictor.
SE(R) is the approximate standard error of the corrected correlation.
W is the factor by which the standard error of the observed correlation is increased due to correction for explicit selection.
N' is the number of additional cases required for the SE(*R*) to equal SE(*r*) when the latter is based on 100 cases.

respectively, the factor by which correcting for range restriction increases the standard error and the number of observations represented by the loss in precision as a result of correcting. The last column, in other words, gives the approximate number of additional observations required for the corrected estimate to be as precise an estimate of the population correlation as the uncorrected estimate is with a sample size of 100.

As can be seen from the entries in the table, under conditions of minimal selection, the loss in precision incurred by correcting for range restriction is relatively minor. When the ratio of standard deviations, *K*, is slightly greater than 1, the standard error of the corrected correlation, SE(*R*), is only 1.065 times larger than the corresponding value for the uncorrected correlation, SE(*r*). However, when the standard deviation in the unselected population is twice as large as it is in the selected sample ($K = 2.0$), SE(*R*) is nearly half again as large as its uncorrected counterpart. In this case, the loss in precision translates into a need for about 117 additional observations in order for a corrected estimate under this degree of selectivity to have as much precision

as has the uncorrected estimate from a sample of 100. When the standard deviation is three times as large in the unselected group, the loss in precision represents a need for 328 additional observations.

Clearly, for situations of extreme degrees of explicit selection on the predictor, the gain in terms of reduced bias via adjusting for selection effects can be completely undermined in terms of increased uncertainty associated with the resulting point estimate (Gross and Kagen, 1983). Moreover, empirical results suggest that this increased uncertainty can have even more deleterious effects under conditions of violated assumptions (Dunbar, 1982). The illustrations in Table 3 reflect approximate sampling fluctuations when specific distributional assumptions are satisfied exactly. If uncertainty also exists with respect to specification of explicit selection variables, the efficiency of a corrected correlation can be expected to decrease even further. This added uncertainty should be kept in mind when interpreting correlations that have been corrected for restrictions of range. Allen and Dunbar (1990) provide more recent results on the sampling behavior of correlations corrected for incidental selection.

Alternative Adjustment Procedures

Recently, the Pearson-Lawley correction formulas have been given a richer conceptual framework by the selection modeling approach to dealing with nonrandom sampling. Although the substantive concerns of researchers using selection modeling techniques are often in the area of evaluation of educational and social programs (e.g. Heckman, 1974; Hausman and Wise, 1976; Cronbach, 1982) with nonequivalent control group designs, their techniques are closely related in intent to the Pearson-Lawley procedures, in that they seek less biased estimates of treatment effects associated with independent variables. As will be seen in the review presented below, many selection modeling approaches derive flexibility in adjusting for sample selection bias at the expense of more stringent assumptions about the nature of the selection process. Some methods also require more complex procedures for the estimation of parameters and larger sample sizes for acceptable degrees of efficiency in parameter estimation. In spite of these difficulties in implementation, a more complete understanding of the validation problems described at the outset of this paper can be gained from a brief review of this general approach.

Muthén and Jöreskog (1983) provide an overview of some basic selection modeling procedures. The bias in r_{xy}, b_o, and b_x when X and Y are incidental selection variables provides a relatively simple case for purposes of illustration. If Z is assumed to be the only explicit selection variable, then the selection process is described by a threshold value, t, such that observations exist in the selected sample only when $Z > t$; in other words, t

is a cut score that truncates the distribution of Z in the selected sample. The regression or conditional expectation of the criterion given values on the predictor of substantive interest then becomes

$$E(Y|X, Z > t) = b_o + b_x X + E(e|Z > t).$$

The final term in the above expression conveys the notion that predictions of one incidental selection variable from another are not unbiased. Here the bias is reflected in a conditional expectation of prediction error that is nonzero. Muthén and Jöreskog (1983) show that $E(e|Z > t)$ depends on the shape of the distribution of Z, the explicit selection variable. When a particular distributional form is specified for Z (or, for that matter, an entire set of Zs), maximum likelihood estimates for parameters in the above equation can be developed giving consistent estimates for the regression parameters and correlation coefficient (cf. Goldberger, 1980). Although particular procedures for accomplishing this have been introduced by Tobin (1958), Heckman (1979), and Muthén and Jöreskog (1983), to mention a few, for the special case of multivariate normality among all variables (explicit and incidental) it also appears that maximum likelihood estimates can be obtained directly from the Pearson and Lawley formulas (Cohen, 1955; Muthén and Jöreskog, 1983). Methods that are based on other distributional forms can require complex iterative procedures for estimating parameters for which convergence to a proper solution is not always guaranteed. Application of these procedures during the normal course of validation activities may not be feasible, given the present state of their development.

In addition to the complete maximum likelihood approaches mentioned above, methods for modeling the selection process as a kind of two-stage procedure have been used with some success (see, for example, Heckman, 1979). In these approaches, the information usually assumed to be available consists of scores on selection variables for all individuals, a binary variable equal to 1 if an individual is selected and 0 otherwise, and a criterion measure for all individuals selected. For multivariate normal selection variables, Heckman's (1979) two-stage approach to adjusting for selection effects involves:

(1) estimating the nonzero conditional expectation of prediction error for individuals via probit regression of the binary variable on the selection variables and
(2) entering this estimate (the selection term) along with the predictors of substantive interest in an ordinary least-squares regression, using criterion data from the selected sample.

The first step in this approach provides an indication, loosely speaking, of the chance that a given individual will be lost due to sample selection;

adding this as a term in the regression equation provides an adjustment for the fact that sample selection is not random. Heckman (1979) showed that this two-stage procedure provides consistent estimates of slopes and intercepts in the regression equation for the selected sample. In validation studies where interest focuses on predicted criterion scores for individuals, as in many studies of differential prediction, the Heckman adjustment can provide less biased assessments of expected performance over a range of predictor scores. This may be relevant to certain types of group comparisons in the Joint-Service context, where a common criterion variable exists. The general approach can also be adapted to provide an alternate means of correcting correlation coefficients, albeit at the expense of distributional assumptions about the selection variables that are not required by the Pearson-Lawley procedures.

A simple example of the effects of the Heckman adjustment illustrates its potential use in criterion-related validity studies. Table 4 shows the results of Heckman's two-stage procedure applied to data from the simulated trivariate normal distribution referred to previously in the context of Figure 4. Recall that the population correlations between Y and each of Z and X were .60, the correlation between Z and X was .90, and the variables were standardized so that the intercept and slope of the Y-on-X regression were 0 and .6, respectively. The table contains adjusted and unadjusted estimates of this intercept and slope for samples selected from either the upper 10 or 50 percent of the population distribution of the selection variable, Z. Also given are the intercept and slope in the sample of 5,000 cases from which the selected samples were actually drawn. Below each parameter estimate the corresponding standard error is given in parentheses.

Entries in Table 4 show the same bias for parameter estimates in the regression equation that was discussed with respect to correlation coefficients. When the selection ratio is .50, the adjustment made by the Heckman procedure is small but provides very accurate point estimates; moreover, the weight assigned to the selection term estimated in the probit stage is moderate. On the other hand, the standard errors of the adjusted values are noticeably larger than those of their unadjusted counterparts, indeed sufficiently larger that the adjusted point estimate of both intercept and slope is within two standard errors of the unadjusted sample value. This phenomenon is exacerbated when the selection ratio drops to .10, with less encouraging results concerning the point estimates themselves. Here the slope is overestimated, the intercept is underestimated, and a band extending two standard errors to either side of the adjusted value encompasses the entire width of the same band around the unadjusted value. Note also that the weights assigned to the selection terms for both selection situations are close enough to zero in standard error units to allow one to incorrectly conclude that selectivity is negligible in each case.

The above results illustrate one feature of the two-stage procedures that

TABLE 4 Illustration of the Heckman Adjustment Under Two Degrees of Selection on Z

Population Correlation Matrix		X	1.0		
		Y	.6	1.0	
		Z	.9	.6	1.0
Regression Equation in Sample of 5,000		*Y* =	-.013 +	.596 (X)	
	Selection Ratio				
	.50			.10	
Parameter	Uncorrected		Corrected	Uncorrected	Corrected
Intercept	.089		-.007	.396	-.340
	(.023)		(.069)	(.108)	(.546)
Slope for *X*	.529		.597	.407	.720
	(.023)		(.053)	(.065)	(.239)
Selection Term	-		.129	-	.314
			(.085)		(.226)

NOTE: Standard errors are given in parentheses.

may limit their utility in the context of test validation. As with any selection modeling approach, the accuracy obtained is contingent on correct specification of the selection process and inclusion of all relevant variables in the probit stage. As noted by Cronbach (1982), if important variables are omitted during estimation of the term representing the probability of selection, then the adjustment will only remove a portion of the bias due to sample selection. This situation parallels the use of Pearson's two-variable correction when selection is actually incidental. However, as more variables are used to describe the selection process accurately, overlap between independent variables in the probit and least-squares stages of the procedure is likely to increase, leading to a selection term that tends to be highly correlated with other predictors in the least-squares model. The resulting collinearity contributes an added degree of instability in the adjusted parameter estimates beyond the instability to be expected simply because adjusted estimates are being employed. In the example, the large standard errors are due in part to the high correlation between *Z* and *X* and to the fact that both variables were included in the probit regression analysis. The problem this poses for criterion-related validity studies lies in the fact that most potential selection variables are precisely those variables whose use as predictors is being validated. Under these circumstances, one might expect the standard errors of the adjusted estimates to be too large to allow for useful inferences

about population correlation coefficients and regression parameters when selection is severe. In other situations, either where selection is moderate or reasonable values for variances in the unselected group are difficult to determine, the two-stage procedure may have some promise.

Some of the technical limitations of the Heckman approach alluded to in the above example have been studied in greater detail, with concern again centering on violated assumptions regarding the distribution of selection variables and on sampling fluctuations of the adjusted estimates. Goldberger (1980) addressed the problem of bias in the adjustment when the selection variables are not normally distributed, suggesting that it can be quite large with only modest departures from normality. When the specific departure is such that the regression of Y on X has a reduced slope in the upper range of the X scores, the two-stage adjustment tends to be conservative, just as did the Pearson-Lawley corrections in this situation (Dunbar, 1982). Unfortunately, this and other selection modeling approaches have seen limited use in empirical studies of criterion-related validity; as a consequence, there is presently little basis for judging how they might perform if used in large-scale validation studies.

Sampling fluctuations and the efficiency of estimates obtained in the selection modeling approaches have also received attention recently, particularly in Nelson's (1984) examination of the behavior of the Heckman two-step method, Olsen's (1980) least-squares version of the method, and a full maximum likelihood version with respect to the degree of collinearity introduced by overlap between variables describing the selection process and variables to be evaluated as predictors. Nelson's results confirm the general suspicions aroused concerning the two-step methods by the example given previously: adjustments for selection bias are most needed precisely when the two-step methods for making them are ineffectual. That is, the high degree of collinearity expected when the selection process is completely specified resulted in unacceptable sampling errors in the parameter estimates from the Heckman and Olsen procedures. Nelson's results showed the full maximum likelihood method to perform better under most circumstances in his simulation study—to the extent that this method is related in spirit to the Pearson-Lawley corrections, some preference for the latter might be inferred from these results. In any case, it appears that a more careful evaluation of some of the alternative approaches to dealing with range restriction problems needs to be made in the context of test validation research before such methods can be used with much confidence on a large scale. Some of the Services are currently engaged in such efforts—results from such studies can have direct implications regarding the expected accuracy of selection modeling approaches in the Joint-Service context (see Rossmeissl and Brandt, 1985; Dunbar, 1986).

A final note on alternative procedures concerns the potential application

of certain Bayesian approaches to handling the missing data problem caused by selection. Rubin (1977), for example, developed a method for assessing the influence of nonrespondents on the results of sample surveys based on subjective notions about the characteristics of nonrespondents. These notions are formalized in terms of the parameters of prior distributions for observations in the nonrespondent population.

In a test validation context, one might consider unselected examinees as analogous to the nonrespondents that Rubin's methods are designed to handle, nonrespondents with particular characteristics because of administrative as well as self selection. Herzog and Rubin (1983) and Glynn, et al. (1986) provide extended discussions and examples of how repeated imputations of missing observations based on a variety of prior distributions can be used together with available data to estimate the desired parameters of a combined population of selected and unselected examinees. The combination of such "mixture models" with repeated imputations from a range of reasonable prior distributions can make explicit the amount of uncertainty that exists regarding the predictive validity of a test. As noted by Rubin (1977), these methods, and perhaps any method for handling selection bias, are best considered as ways of formalizing the possible effects of missing observations on outcome measures rather than as substitutes for random samples normally required for valid inferences about the characteristics of a population.

IMPLICATIONS FOR PREDICTIVE VALIDITY IN A JOINT-SERVICE CONTEXT

The problem of range restriction is not new to anyone concerned with establishing the criterion-related validity of selection and classification tests in the Services, although the methods for coping with it have been far from uniform over the years. Some Services have used either the Pearson or Lawley corrections routinely in reporting the results of validity studies for the ASVAB and its predecessors, while others have questioned this use, especially when validation is conducted within particular occupational categories (cf. Sims and Hiatt, 1981; and Air Force Human Resources Laboratory, 1982, for an instance of this contrast of viewpoints). Because many things can affect the magnitudes of correction factors, comparison of corrected correlations across Services is not a simple task. The choice of analytical procedures, base or reference populations to which the corrected estimates are intended to generalize, and the variables that are to serve as explicit selection measures can all influence the magnitudes of corrected estimates of predictive validity. These issues are discussed in what follows as they relate to the use of adjustment procedures for validating the ASVAB against measures of performance in current use for military jobs, and for

validating alternative measures of job proficiency (surrogate criteria) against on-site evaluations of job performance (benchmark measures).

Analytical Procedure

The demands on a procedure for dealing with selection effects in the Joint-Service context for military performance assessment are not typical of many settings in which test validation studies are conducted. One purpose behind Service-wide efforts in this regard is the achievement of a degree of comparability in characterizations of predictive validity

(1) across Services,
(2) across jobs of a similar nature within and between Services, and
(3) across jobs that involve different tasks and hence different combinations of ASVAB subtests for selection.

Moreover, a concern exists for examining the consistency of predictions of performance criteria for groups of military trainees distinguished by sex, race or ethnicity, and level of education. Comparisons of this kind, which make the cooperative venture especially useful, can be hopelessly confounded by varying degrees of range restriction in the groups involved. This problem is duly noted with respect to comparisons across occupations in the work of Schmidt and Hunter (1977) on validity generalization, and with respect to comparisons across demographic groups in the work of Linn (1983a, 1983b) on differential prediction. An important observation regarding the purpose of corrections for range restriction with such comparisons in mind is that they are as much needed to obtain comparability as they are to provide precise estimates of population values. Because they are also to be used with a variety of criterion measures (surrogate as well as benchmark measures of job performance), their limitations need to be fully understood and appreciated.

As indicated at various points in the preceding review, the standard Pearson-Lawley correction procedures are familiar to most personnel psychologists and appear to have been carefully evaluated in both analytical and empirical studies, many of the latter being performed with the specific problems of criterion-related validity in mind. The limitations of these procedures and the conditions under which they are likely to give misleading indications of predictive validity are well documented. In contrast to these conventional techniques, the procedures based on selection modeling are in comparative infancy. Their relationship to the conventional procedures is only partially understood and their use in the context of predictive validity studies has been all but nonexistent. In addition, the sample size requirements of the more sophisticated estimation methods accompanying some of the selection

modeling procedures may not be met by many military job classifications for which validity checks are desired. Thus, it is probably premature to endorse these newer methods in the context of ASVAB validation. Regarding their use on a routine basis, more caution is probably warranted due to the multiplicity of alternate methods and to problems in the technical implementation of some. It is also premature, however, to dismiss these methods as inappropriate for criterion-related validity studies in general—further investigations of these techniques with military performance data is to be encouraged.

Reference Population

Results from the application of any adjustment procedure necessarily reflect the characteristics of the group from which information about the unselected population is obtained. In cases where the Pearson-Lawley corrections are used, the source of estimates of the variances and covariances of selection variables in the unrestricted group, in addition to the selection process itself, determines the magnitude of any correction factor. Variations in these estimates, such as those caused by preexisting differences between potential enlistees opting for one Service over the others, result in correction factors of varying magnitudes and make the corrected values difficult to interpret. For these reasons, when comparison across Services is important, using the entire accession populations within Services in a given year as reference groups would be counterproductive. The corrected predictive validities for selection composites would differ from Service to Service as well as from year to year. Clearly, a reference population common to all Services reporting predictive validities for the ASVAB is desirable.

Recent versions of the ASVAB have been anchored to a nationally representative sample of men and women drawn as part of the National Longitudinal Survey (NLS) of Youth Labor Force Behavior, sponsored by the Departments of Labor and Defense and usually referred to as the 1980 Youth Population (U.S. Department of Defense, 1982). Although the 1980 Youth Population does not precisely reflect current applicant pools for each Service, it does provide a frame of reference that would allow corrected correlations to be directly comparable across Services. Moreover, it does not constitute a group about which concerns over self-selection phenomena would arise, as would be the case for an accession population.

In spite of the fact that the 1980 Youth Population is the closest example of any kind of normative group for dealing with the effects of range restriction, there are some limitations in using it. As a representative sample of the nation's youth, this group contains individuals who would be judged ineligible for military service on the basis of ASVAB scores used in the initial screening of enlistees and is thus atypical of a projected mobilization

population in the wider range of talent represented. With this wide range of talent, one could expect correction factors to be quite large in some cases, certainly larger than expected when interest is focused on the predictive validity of ASVAB composites for only those recruits who have passed an initial screening. In view of the fact that corrections based on the total 1980 Youth Population might be artifactually large, some portion of this group might be best chosen to meet the need of adjusting validity coefficients for comparability across Services. The effects of suggested restrictions of the Youth Population on the magnitudes of correction factors are illustrated in the example presented below.

In order to provide an idea of the kinds of results to be expected when using the Pearson-Lawley correction procedures in connection with the 1980 Youth Population, an example is given in Table 5 based on training data from nine Marine Corps clerical specialties. Given in Table 5 are uncorrected and corrected correlations between the ASVAB clerical composite from Forms 8/9/10 used by the Marine Corps and final course grades in training. The Lawley corrections given in the table assume the 10 individual subtests to be explicit selection variables, and the composite and course grades to be incidental selection variables. The adjusted values given in column A were obtained by using the variances and covariances of subtest standard scores for the total Youth Population (U.S. Department of Defense, 1982). To approximate the situation in which only that portion of the Youth Population eligible for military service is used, subtest standard deviations from the total sample were adjusted for four degrees of truncation under the assumption that each subtest followed a normal distribution. The resulting reduced standard deviations were used along with the original correlations from the total sample in obtaining a variance-covariance matrix among explicit selection measures that would provide a rough indication of how much smaller correction factors might be with ineligible examinees deleted from the reference group. This procedure probably underestimates the amount by which correction factors would change if, for example, the bottom 10 percent of the AFQT distribution (the actual screening composite) were deleted from the reference group since only the standard deviations were altered in the calculations.

As can be seen from the entries in Table 5, the Lawley corrections suggest that the amount of range restriction in these groups is substantial in nearly all cases. For the specialties with reasonably large sample sizes, the smallest difference between corrected and uncorrected validity coefficients is .19 in the second administrative group. For other training courses, the corrections in column A based on the entire Youth Population are quite large. The range restriction phenomenon is further illustrated in the lower portion of Table 5, which contains subtest standard deviations for each training cohort and for the entire Youth Population. Inspection of these

TABLE 5 Corrected Predictive Validities of the Marine Corps Clerical Composite in Nine Clerical Training Programs Based on 1980 Youth Population and Subtest Standard Deviations in Selected and Unselected Groups

Panel (a)

Training Program	*N*	Uncorrected Validity	Corrected Validities A	B	C	D	E
Administrative 1	632	.32	.56	.52	.50	.48	.46
Administrative 2	620	.25	.44	.41	.39	.37	.36
Communications center	332	.19	.56	.52	.50	.48	.47
Supply stock	653	.47	.69	.65	.63	.62	.60
Aviation supply	379	.30	.55	.51	.49	.47	.45
Finance records	227	.20	.65	.62	.60	.58	.57
Basic preservation	51	.28	.36	.34	.33	.32	.31
Subsistence supply	66	.11	.59	.56	.54	.52	.51
Aviation operations	87	.36	.66	.63	.61	.59	.57

Panel (b)

Training Program	AR	WK	PC	NO	GS	CS	AS	MK	MC	EI
Administrative 1	6.95	6.25	7.43	8.11	7.21	9.45	8.13	7.09	8.34	7.87
Administrative 2	7.05	6.34	7.21	6.82	7.57	9.46	8.32	7.57	8.17	7.51
Communications center	7.91	7.12	8.49	6.53	8.55	9.51	8.96	7.58	8.77	8.55
Supply stock	7.19	5.72	6.96	6.35	7.39	8.37	8.81	8.03	8.27	7.96
Aviation supply	7.21	6.16	8.17	8.31	7.13	9.08	8.36	7.34	8.34	7.76
Finance records	5.65	3.94	5.10	5.73	6.69	7.78	8.37	6.81	7.37	7.89
Basic preservation	8.05	7.93	8.98	7.13	7.95	6.46	8.12	8.13	8.31	8.59
Subsistence supply	7.05	6.87	7.14	5.75	7.48	8.13	8.00	7.03	7.98	7.98
Aviation operations	6.92	6.20	7.73	7.30	6.35	8.53	8.26	6.77	8.53	8.09
1980 Youth Population	10.25	10.05	9.66	10.65	9.69	10.10	9.92	10.77	9.55	9.86

NOTE: Corrections using the Lawley formula make use of standard deviations on 10 ASVAB subtests from Form 8, derived from:

A the total 1980 sample.
B the upper 95 percent of the 1980 sample.
C the upper 90 percent of the 1980 sample.
D the upper 85 percent of the 1980 sample.
E the upper 80 percent of the 1980 sample.

values shows the varying degrees of selectivity across groups. Generally speaking, the standard deviations in the training cohorts are one-half to two-thirds the size of the corresponding values in the proposed reference population.

The effects of removing ineligible individuals from the Youth Population prior to adjusting the predictive validities of the clerical composite for selection

effects are shown in columns B through E in the upper portion of Table 5. Even after deleting lower scoring individuals, it appears that corrected values remain likely to be substantially larger than the uncorrected ones in some cases—extreme selection in some occupational specialties, such as the financial records clerk group in this instance, makes this result unavoidable. However, the relative magnitudes of the corrected validity coefficients across training programs remain the same regardless of the proportion of the reference group that is deleted.

It should be noted that the regular pattern of steadily decreasing estimates of predictive validity for increasing degrees of truncation in the proposed reference population is to be expected when the selection variables (here ASVAB subtests) have symmetric distributions, as assumed in this illustration. ASVAB subtests are well known to have nonsymmetric distributions. This is likely to influence the pattern of decreasing estimates of validity when the reference population is truncated on AFQT. Indeed, Maier (1985) has shown that the effects of truncation on the multivariate corrections are more complex than the example here would indicate, in part because of skewness in the distributions of selection variables and in part because of the effects of truncation on the covariances of selection variables. Neither factor is considered in the example presented above.

Selection Variables

In the above example, individual ASVAB subtests served as explicit selection variables. In practice, most job training programs in the military make use of one or more composite measures in selecting recruits for training. Typically, cut scores for selection are established on the composite scale separately for high school graduates and nongraduates, making high school graduation an additional selection variable. This custom poses several problems for the usual correction procedures in the Joint-Service context. First of all, ASVAB selection composites are not uniform from Service to Service, so that using the actual selection measure would again militate against the goal of comparability in the resulting values. Here, comparability is in direct conflict with accurate specification of the selection process. If using the subtests instead of the composite measures means that the selection mechanism is incompletely specified, the Pearson-Lawley corrections might likely give conservative indications of the predictive validity of the composites. Second, the inclusion of graduation status as a dichotomous selection variable in the Pearson-Lawley formulation is reasonable in principle, but complicated by the fact that certain job training programs admit only recipients of high school diplomas. No variability in the selected sample for such groups represents a kind of limiting case for the Pearson-Lawley procedure, in which no simple correction for range restriction is feasible. It would therefore seem appropriate to consider ASVAB subtests as the only available

measures that are both closely related to the selection process and common to all Services. Any systematic errors incurred as a result of this choice are likely to result in corrected validities that are smaller than they might otherwise be with more detailed specification of the selection mechanisms operating for individual training and occupational cohorts.

In considering the choice of variables used in correcting for range restriction, the properties of criterion measures should not be overlooked. Selection effects are likely to increase when surrogate and benchmark measures are used as criteria. General attrition can be expected to alter the distribution of both types of measures relative to the distribution of final grades in training, and the logistic problems of collecting systematic on-site evaluations of performance could have unknown effects on the range of talent represented in the distributions of benchmark variables. If selectivity is augmented by factors such as these, both the bias error and sampling error of corrected correlations will be affected and the degree of uncertainty regarding the resulting values magnified. The possible presence of additional sources of unreliability in alternative performance measures is another area of concern. While it may be the case that all of these factors (i.e., added selectivity and measurement error) suggest an increased tendency for the Pearson-Lawley corrections to be conservative, their exact influence in individual cases is difficult to determine. To the extent that unreliability may effect corrections that are overly conservative, additional corrections for attenuation might be appropriate. These would of course increase the sampling errors of observed values even further.

CONCLUDING REMARKS

The effects of selection on correlation coefficients and regression parameters place the personnel psychologist on the horns of a classic statistical dilemma. To retain observed values gives an extremely misleading view of the relationship between predictor and criterion variables, but to correct observed values places one at the mercy of assumptions that will not be strictly satisfied in practice and of an added degree of uncertainty in estimating population values. While there is no inherent magic in any procedure for dealing with the effects of selection bias, neither is there an inherent sorcery. Rather, a balanced indication of the quality of a given predictor can be achieved by reporting both uncorrected and corrected validity coefficients and by careful documentation of the methods used to obtain the corrected estimates. The tenability of assumptions can be examined in individual cases through the use of graphical techniques described elsewhere, and in some cases reasonable speculation about the influence of violated assumptions can be entertained. The review of analytical techniques suggested the use of the standard Pearson-Lawley correction formulas in validation

studies involving Service-wide applications. A common reference population and set of explicit selection variables also seem desirable for any degree of comparability to be achieved through corrections for restriction of range. When a summary statistic is needed to describe the predictive validity of a selection instrument in a variety of settings, these approaches can provide a more complete assessment of the relationship between that instrument and the relevant measures of on-the-job performance.

REFERENCES

Air Force Human Resources Laboratory 1982 Aptitude Index Validation of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 5, 6, and 7. Air Force Human Resources Laboratory, Personnel Research Division, Lackland Air Force Base, Tex.

Allen, N.L., and S.B. Dunbar 1990 Standard error of correlations adjusted for incidental selection. *Applied Psychological Measurement*, 14:83-94.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland 1975 *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass.: MIT Press.

Bobko, P., and A. Rieck 1980 Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement* 4:385-398.

Booth-Kewley, S. 1985 An Empirical Comparison of the Accuracy of Univariate and Multivariate Corrections for Range Restriction. TR-85-19. Navy Personnel Research and Development Center, San Diego, Calif.

Brandt, D., D. McLaughlin, L. Wise, and P. Rossmeissl 1984 Complex Cross-Validation of the Validity of a Predictor Battery. Paper presented at the annual meeting of the American Psychological Association, Toronto.

Brewer, J., and J.R. Hills 1969 Univariate selection: the effects of size of correlation, degree of skew, and degree of restriction. *Psychometrika* 34:347-361.

Cohen, A.C. 1955 Restriction and selection in samples from bivariate normal distributions. *Journal of the American Statistical Association* 50:884-893.

Cronbach, L.J. 1982 *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Cronbach, L.J., D.R. Rogosa, R.E. Floden, and G.G. Price 1977 Analysis of Covariance in Non-Randomized Experiments: Parameters Affecting Bias. Occasional paper, Stanford Evaluation Consortium, Stanford University.

Dunbar, S.B. 1982 Corrections for Sample Selection Bias. Unpublished doctoral dissertation, Department of Educational Psychology, University of Illinois at Urbana-Champagne.

1986 Comparing Regression Equations Across Training Programs: An Empirical Study of Prior Selection Effects and Alternative Predictive Composites . ONR Technical Report 86-1: Measurement Research Group, Iowa City, Iowa.

Forsyth, R.A. 1971 An empirical note on correlation coefficients corrected for restriction of range. *Educational and Psychological Measurement* 31:115-123.

Glynn, R.J., N.M. Laird, and D.B. Rubin 1986 Selection modeling versus mixture modeling with non-ignorable nonresponse. In H. Wainer, ed., *Drawing Inferences from Self-Selected Samples*. New York: Springer-Verlag.

Goldberger, A.S. 1980 Abnormal Selection Bias. Workshop Paper 8006. Social Systems Research Institute, University of Wisconsin-Madison.

Greener. J.M., and H.G. Osburn 1979 An empirical study of the accuracy of corrections for restriction in range due to explicit selection. *Applied Psychological Measurement* 3:31-41.

1980 Accuracy of corrections for restriction in range due to explicit selection in heteroscedastic and non-linear distributions. *Educational and Psychological Measurement* 40:337-346.

Greenlees, J.S., W.S. Reece, and K.D. Zieschang 1982 Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association* 77:251-261.

Gross, A.L., and L. Fleischmann 1983 Restriction of range corrections when both distribution and selection assumptions are violated. *Applied Psychological Measurement* 7:227-237.

Gross, A.L., and E. Kagen 1983 Not correcting for restriction of range can be advantageous. *Educational and Psychological Measurement* 43:389-396.

Gullickson, A., and K. Hopkins 1976 Interval estimation of correlation coefficients corrected for restriction of range. *Educational and Psychological Measurement* 36:9-25.

Gulliksen, H. 1950 *Theory of Mental Tests*. New York: John Wiley & Sons.

Hausman, J.A., and D.A. Wise 1976 The evaluation of results from truncated samples: the New Jersey income maintenance experiment. *Annals of Economic and Social Measurement* 5:421-445.

Heckman, J.J. 1974 Shadow prices, market wages, and labor supply. *Econometrica* 42:679-694.

1979 Sample selection bias as a specification error. *Econometrica* 47:153-161.

Herzog, T.N., and D.B. Rubin 1983 Using multiple imputations to handle nonresponse in sample surveys. In W.G. Madow, I. Olkin, and D.B. Rubin, eds., *Incomplete Data in Sample Surveys*, *Volume 2*. *Theory and Bibliographies*. New York: Academic Press.

Kelley, T.L. 1923 *Statistical Method*. New York: MacMillan & Co.

Lawley, D. 1943A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh:*

1944 *Proceedings*, *Section A* 62:28-30.

Linn, R.L. 1968 Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin* 69:69-73.

1983a The Pearson correction formulas: implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement* 20:1-15.

1983b Predictive bias as an artifact of selection procedures. In H. Wainer and S. Messick, eds., *Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Linn, R.L., D.L. Harnisch, and S.B. Dunbar 1981 Corrections for range restriction: an empirical investigation of conditions resulting in conservative corrections. Journal of Applied Psychology 66:655-663.

Lord, F.M., and M.R. Novick 1968 *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.

Maier, M.H. 1985 Effects of Truncating a Reference Population on Corrections of Validity Coefficients for Range Restriction. Research Memorandum 85-40. Center for Naval Analyses, Alexandria, Va.

Muthén, B.. and K.G. Jöreskog 1983 Selectivity problems in quasi-experimental studies. *Evaluation Review* 7:139-174.

Nelson, F.D. 1984 Efficiency of the two-step estimator for models with endogenous sample selection. *Journal of Econometrics* 24:181-196.

Novick, M.R., and D.T. Thayer 1969 *An Investigation of the Accuracy of the Pearson Correction Formulas* . Research Memorandum 69-22. Princeton, N.J.: Educational Testing Service.

Olsen, R.J. 1980 A least-squares correction for selectivity bias. *Econometrica* 48:1815-1820.

Pearson, K. 1903 Mathematical contributions to the theory of evolution—XI: On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society*, *London* , *Series A* 200:1-66.

Rossmeissl, P., and D. Brandt 1985 Modeling the Selection Process to Adjust for Restriction in Range. Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Rubin, D.B. 1977 Formalizing subjective notions about the effects of nonrespondents in sample surveys. *Journal of the American Statistical Association* 72:538-543.

Rydberg, S. 1963 *Bias in Selection*. Stockholm: Almquist and Wiksell.

Schmidt, F.L., and J.E. Hunter 1977 Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology* 62:529-540.

Sims, W.H., and C.M. Hiatt 1981 Validation of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7 with Applications to ASVAB Forms 8, 9 and 10. Report 1160. Center for Naval Analyses, Alexandria, Va.

Thorndike, R.L. 1949 *Personnel Selection*. New York: John Wiley and Sons.

Tobin, J. 1958 Estimation of relationships for limited dependent variables. *Econometrica* 26:24-36.

U.S. Department of Defense 1982 Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery. Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics), Washington, D.C.

Alternatives to the Validity Coefficient for Reporting the Test-Criterion Relationship

Linda J. Allred

Test scores are widely used as the basis for personnel decisions such as selection and placement. Applicants for clerical positions must pass a typing test, colleges require some minimum score on the Scholastic Aptitude Test, and government employees take the Civil Service examinations. Even preschoolers must meet minimum standards on intelligence tests for admission to many private kindergartens. These testing programs provide an objective method of screening individuals to find those best qualified for the job in question. (The use of the word job in this sense refers to any employment, training, or academic situation in which a testing program is used.)

The effectiveness of a testing program in selecting appropriate individuals depends on how well the test scores correspond with some objective measure of actual performance. The relationship of test scores to an objective performance measure is called predictive validity.

The assessment of validity requires first of all a performance measure to serve as the criterion against which the test is validated. The criterion might be simply success or failure, or retention or dismissal, but a more precise measure of achievement or performance allows a better validation. Here, a suitable measure of performance is presumed to exist. In addition, for purposes of this discussion, it is assumed that a group of persons is tested and then hired, selected, or appointed without regard to the test, so that the full potential effect of the test can be examined.

The relationship between test scores and performance is often expressed

as a validity coefficient (r), a number between 0 and 1.0 that indicates how well the performance measure, or criterion, is predicted by the test. The validity coefficient indicates the overall strength of the test-criterion relationship for the group being studied, but its meaning is obscure to a nontechnical audience. Nor does it help much to say that r^2 (the squared validity coefficient) indicates the proportion of performance variance accounted for by the test. Other more direct means are needed for displaying the meaning of a validity correlation coefficient. Moreover, the coefficient does not provide full information about this relationship.

Various methods are available for displaying detailed information about the test-criterion relationship. These methods range from a full plot of test and performance score data to ways of grouping test scores into intervals and then providing detailed information about the distribution of performance scores for groups of individuals in each test score interval.

Several sets of hypothetical data are shown here to illustrate these display methods. Each data set includes a test score and a performance score for 1,000 individuals. To simplify the interpretation of these data sets, both the test and the performance measure range from a minimum of 0 to a maximum of 50.

The simplest method of displaying the test-criterion relationship is the scatter plot. This is a graph showing each individual's scores on the test and the performance measure. The test score is normally listed on the horizontal axis and performance is listed on the vertical axis. Figure 1a shows the scatter plot for data set A (all figures in this chapter are located at the end of the text). Each * represents a single individual, and a number indicates that more than one individual had that test and performance score. For example, point a represents an individual with a test score of 11 and a performance score of 22. This data set illustrates a very strong relationship between test score and performance. The individuals in this group who score high on the test also score high on the performance measure, and vice versa. The validity coefficient for this data set is .96.

The difference between a strong and a moderate relationship is apparent if Figure 1a is compared to Figure 1b, which shows the scatter plot for data set B. In this example, there is still a tendency for the highest test scores to be associated with high performance scores and vice versa, but there is much more variation than was seen with data set A in Figure 1a. The validity coefficient for data set B is .52.

Figure 1c shows a very weak relationship. There is a wide range of scores on the performance measure for each possible test score. For example, individuals who have a test score of 3 have performance scores ranging from 0-50, while those with a test score of 43 also range from 0-50 on the performance measure. Thus, there is relatively little difference in performance at high and low test scores. The validity coefficient for this data set is only .19.

Comparison of these three scatter plots gives an indication of the differences in data distribution for tests with different validities. In addition, the scatter plot is useful for detecting differences in distribution among tests with the same validity. Figures 1d and 1e show scatter plots for two distributions, both with a validity coefficient of .65. Although the tests have the same validity, it is apparent from comparison of the scatter plots that the two distributions are very different. Reporting the validity coefficient alone for either of these data sets would not reveal the abnormality of the distribution.

Data sets D and E illustrate a special kind of validity problem that exists when the predictive ability of the test is not equal at all levels of the test. In Figure 1d (data set D), high test scores are associated with only high performance scores, yet low test scores are associated with the full range of performance scores. In Figure 1e (data set E) the opposite is true, with less variability in performance scores at lower test score levels. This type of situation can occur for a number of reasons. Aptitude tests, for example, may measure prior experience with the subject matter, rather than ability; individuals with prior experience do well on the test and on the performance measure. However, some individuals who do poorly on the test will gain experience on the job and also do well on the performance measure. Similar problems are common in diagnostic testing when a positive test result confirms the presence of a condition but a negative result does not rule out the condition, or vice versa.

The scatter plot is an important tool in evaluating test validity, but its usefulness tends to be limited to giving the evaluator a general idea of the regularity of the distribution. In order to get more specific information about the test-criterion relationship, it is important to look at some measure of representative criterion performance at various test score levels. The most obvious way to do this is to plot an average performance score for each test score interval.

The purpose of testing is to predict performance from a test score. Without the test, the best prediction of performance that can be made is the average performance score for the entire group. For example, if we know only that the average grade point average for college freshmen is 2.5, and we know nothing else about a particular high school senior, then the only prediction we can make about that senior's performance the following year is the group average, or 2.5. However, we will be making the same prediction for all students, so predicted performance will not allow us to distinguish between students who will do well and those who will fail. On the other hand, if we know that this student has a score of 98 on a college entrance examination and that the average grade point average for students who score between 95 and 100 on that entrance exam is 3.5, then we can feel more confident that this particular student will do well in college.

In Figure 2, average performance is plotted at 5-point intervals of test

score for data sets A, B, and C. (In this particular example, the steepest line also represents the most valid test, but this is not always true. The steepness of the line depends on the scale of both test and criterion. Relative steepness indicates relative validity only when comparing test scores and criterion measures of similar scale.) For data set A, the predicted performance score for individuals with test scores in the 0-5 interval would be 3.3, while the prediction for individuals in the 46-50 interval would be 46.5. For the least valid test, data set C, there is relatively little difference in predicted performance score between test score intervals. However, displaying only the arithmetic average, or mean criterion score, disregards the variability of these performance scores about the mean.

For example, Figure 3a shows a simple plot of mean performance score for five intervals of a test score. This plot could represent any one of the situations in Figures 3b-e. In Figure 3b, there is a perfect relationship between test score and performance; knowing an individual's test score would permit a perfectly accurate prediction of performance. In Figure 3c, there is a small amount of variability around the mean performance score, so that the prediction of performance from a test score will not always be perfect, but it will be very nearly so. Figure 3d shows substantial variability; performance scores at one test score interval considerably overlap those at adjacent intervals. Prediction of performance scores at each test score interval will involve much error. The distribution in Figure 3e shows much variability at lower test score intervals, decreasing to no variability at the highest interval (similar to the situation in Figure 1e). Predicted performance at high test score intervals will be very accurate, but there will be error in prediction at lower intervals.

The implications for personnel decisions based on the plot in Figure 3a would vary widely depending on which distribution is involved. Unless some indication of variability is also included, however, it would be impossible to tell which distribution is represented. For this reason, the plot of means alone may be misleading.

The box-and-whisker plot in Figure 4 provides a method of displaying both a measure of representative performance and variability about that measure. For each interval of test score, a line (*a*) is drawn between the lowest and highest values of performance. This is the "whisker" and represents range. The "box" is formed by placing a short line (*b*) perpendicular to the range line at the point of representative performance and two lines (*c* and *d*) on either side of that line to indicate variability. The lines are then connected, forming a box. Two types of box-and-whisker plot are generally used. In the first, the measure of representative performance is the arithmetic average, or mean. Variability is the standard deviation, which is an index of the differences between individual scores and the mean. In a normal, symmetrical distribution, 68 percent of the individual scores will be

within one standard deviation above and below the mean. Since the validity coefficient is based on the mean and standard deviation, this method provides a direct display of the information contained in the validity coefficient.

Figures 5a-e show box-and-whisker plots for data sets A-E using the mean and standard deviation. By comparing each plot to the corresponding scatter plot in Figure 1, it is possible to see that the box-and-whisker plot provides a convenient summarization of the information in the scatter plot. For an individual in any given test score interval, the best prediction of performance is the most representative score, in this case the mean. The size of the box indicates how tightly clustered performance scores are about the mean. When the range is narrow and the box is small, as in the test score intervals in Figure 5a (data set A), then the prediction of performance will be very accurate.

In Figure 5b, the moderate validity coefficient is reflected in the longer whiskers and boxes in the plot. Performance scores are less tightly clustered about the mean in each test score interval, so there is more vertical overlap among intervals, and the range differs less among intervals than with the plot of data set A in Figure 5a. As a result, there will be more inaccuracy in the prediction of performance score from test score (so the validity coefficient is lower). Figure 5c shows the weak test-criterion relationship for data set C. If the box-and-whisker plot is compared to the scatter plot for this data set in Figure 1c, it is apparent that the box-and-whisker plot is much clearer. At almost every test interval, the full range of performance scores is represented and the boxes are very long, indicating that performance scores are distributed broadly about the mean. If the best prediction, the mean, is used, there will be considerable error in predicted performance.

The abnormalities in the distributions for data sets D and E are clear in the box-and-whisker plots in Figures 5d and 5e. It is now easy to see the areas of the distributions in which there will be the greatest prediction error. For data set D (Figure 5d), prediction will be fairly accurate at high test score intervals, but there will be significant error in the prediction of performance for those with lower test scores. For data set E, prediction will be more accurate at lower test scores, with many errors at the upper test score intervals.

For specific types of personnel decisions, this information is critically important. If the cost of poor performance is very high, then a selection test needs to be very accurate in selecting applicants who will not perform poorly. The test in data set D will permit selection of only those applicants who will do well, while the test in data set E would not. For data set E, at any test score interval many individuals will be selected who will perform poorly. The box-and-whisker plot makes it possible to see exactly where the most prediction error will occur with the test in question.

In addition to the mean and standard deviation, percentiles can be used as the measures of representative performance and variability in the box-and-whisker plot. Because percentiles are more readily acceptable to the nonstatistician, this method may often be preferable. The 50th percentile, or median, is used instead of the mean, and the 25th and 75th percentiles are used instead of the standard deviation to form the box. The whisker still indicates the range from the 1st to the 100th percentiles.

Figures 6a-e show box-and-whisker plots based on percentiles for data sets A-E. In the first interval of data set A (Figure 6a), the 50th percentile is 3. Twenty-five percent had performance scores higher than 5 (the 75th percentile), and 25 percent were at 1 or below. As can readily be seen by comparing these figures to Figures 5a-e, the information is very similar but is presented in terms more commonly understood. In addition, unlike the mean and standard deviation, percentiles are less affected by outliers—single individuals who score much higher or lower than others in the group.

In the examples above, test scores were divided into 10 five-point intervals. By using equal intervals, the general shape of the original distribution is maintained. However, it is possible to use other methods of dividing the test scores. Percentiles can be used as the basis for dividing test scores, so that the first interval is composed of the lowest 10 percent of test scores, the second interval is the next 10 percent, etc. This will result in approximately equal numbers of individuals in each interval, although the shape of the distribution may be somewhat distorted. In addition, standard scores, such as T-scores, may be used for determining the intervals, but these standard scores often require some degree of expertise to interpret.

In many test-criterion situations, some minimal level of acceptable performance is present. For example, most colleges have a minimum grade point average that must be maintained for an individual to remain enrolled. Performance scores above the minimum or cutoff are considered successes, while those below are considered failures. In this situation, an important validity issue is not the prediction of a performance score per se, but rather the prediction of success or failure. The expectancy chart displays information about successes or failures at different test score intervals. In general, test score intervals are listed in the left column, and bars are used to represent the proportion of successes or failures in each interval. As with the box-and-whiskers plot, test score intervals may be defined as equal intervals of raw test scores, as percentiles, or as standard scores.

The expectancy chart provides a simple method of evaluating the utility of a test. In most test-criterion situations, test users must consider the numbers of successful and failing individuals in economic terms, i.e., the relative cost of success and failure must be considered. The strategy used in implementing a test will depend on the test user's needs. In some cases, a specific number of individuals need to be selected. If the cost of failure is

high, then the test user will need to minimize the number of applicants selected who fail. On the other hand, if the cost of failure is low, then many applicants can be selected. The expectancy chart can be used to evaluate the efficiency of different test score cutoffs to best serve the test user's specific needs.

To illustrate the expectancy chart in Figures 7a-e, an arbitrary criterion cutoff of 25 has been applied to samples A-E. Success is thus defined as a performance score greater than 25, and failure is a score of 25 or below. The proportions of successful individuals are given at five-point test score intervals for these data sets. For data set A (Figure 7a), 100 percent of the individuals with test scores in the 46-50 interval succeed, dropping to 0 percent at the 6-10 score interval. It would be possible with this test to establish a test score cutoff so that no individuals who are selected fail. With a test of high validity, the proportion of individuals who succeed should be very high at the upper intervals, dropping abruptly to zero as the performance cutoff is reached.

By comparison, for data set B (Figure 7b), 79 percent of the highest test score group and 14 percent of the lowest test score group succeed. At any test score cutoff, some individuals will fail, although the proportion will be much lower if the test score cutoff is relatively high. For data set C (Figure 7c), test efficiency is clearly low. The proportion of individuals who will succeed if selected is only 67 percent at the highest test score interval and drops only to 41 percent at the lowest interval. Thus, for any test score interval, there will be almost equal numbers of successes and failures.

The differences between the distributions for data sets D and E are also apparent in the expectancy charts for these data sets (Figures 7d and 7e). Although these distributions have the same validity coefficient, the greater efficiency of the test in data set D for selecting successful individuals is apparent in Figure 7d. One hundred percent of the individuals in each of the upper four test score intervals succeed, while the proportions of successes at the upper intervals in Figure 7e are much lower.

In addition to reporting either the proportion of successes or failures within each test score interval, it is possible to use the expectancy chart to display the proportion of total individuals selected who would succeed at various test score cutoffs. This type of expectancy chart is useful for determining what cutting score would best serve the test user's needs. Figures 8a-e show the proportion of total individuals selected who would succeed at test score cutoffs in steps of five. The efficiency of tests with different validities is evident when the expectancy charts for data sets A-C are compared.

With data set A, (Figure 8a) at a cutoff score of 36 on the test 100 percent of the individuals selected will succeed, dropping to only 51 percent if no cutting score is used. For B, (Figure 8b) at any cutting score the

proportion of successes is lower, ranging from 78 percent at the highest cutting score to 55 percent if no cut is used. With C, (Figure 8c) the maximum proportion of successes is only 67 percent. Once again, the relative efficiency of the tests in samples D and E is clear (Figure 8d and 8e).

The information in the expectancy chart can also be demonstrated by plotting the proportion of successes in each test score interval or at each cutting score. In the expectancy plot, the proportion of successes in each interval gives a clear indication of how well a test discriminates among individuals. If a test is a perfect predictor of performance, then it is possible to find a test score above which all individuals succeed and below which all individuals fail. The plot of such a test is shown in Figure 9a. At the five lowest test score intervals, all individuals fail, while all individuals in the upper five test score intervals succeed. In contrast, Figure 9b shows a test that does not discriminate among individuals. At each test score interval, the proportion of successes is 50 percent.

In Figures 10a-e, the proportion of successes is plotted at 10 five-point test score intervals for the sample data sets. By comparing these plots to the plots in Figures 9a and 9b, it is apparent that the test in sample provides the best overall discrimination between successes and failures. This type of plot is particularly useful for comparing different tests for use with the same performance measure.

In addition, it is often important to determine how much improvement in the proportion of successes occurs at different cutting scores on the test. When no cutting score is used, that is, when the test is not used, the proportion of individuals who succeed is called the base rate. By plotting the proportion of successes at each cutting score, it is possible to evaluate the use of a test compared to the base rate alone. Figures 11a-e show the proportion of successes at ten cutting scores for the five sample data sets. The relative efficiency of the test in data set A (Figure 11a) is displayed by the fairly sharp climb from the base rate (51 percent successes) to 100 percent. With the moderate validity in data set B (Figure 11b) the climb is less steep and 100 percent is not reached, while with the low validity in data set C there is little improvement in the proportion of successes across all cutting scores.

An important issue in many testing situations is not only how many individuals selected by the test succeed but also how many of those rejected would not have succeeded. In Figure 12 the proportions of individuals not selected who would have failed is superimposed on the plots of proportions of successful individuals from Figure 11. This type of plot is useful for determining the most efficient cutting score. For example, in Figure 12a (for data set A), at a cutting score of about 26 on the test, over 90 percent of those selected would succeed, while over 90 percent of those not selected would have failed. The abnormal distributions in data sets D and E are

reflected in the plots in Figures 12d and 12e. For data set D, at the higher cutting scores all of those selected succeed, but many individuals not selected would have succeeded also.

The expectancy table also provides a method of displaying information about the proportion of individuals at various test and criterion levels. It is possible to examine the proportion of individuals in each test score interval who perform at various criterion levels. Figure 13 shows the proportion of individuals in five 10-point test score intervals with performance scores in each of five 10-point intervals. In sample A (Figure 13a), 87 percent of the 152 individuals with test scores of 10 or below also have performance scores of 10 or below, and 13 percent have performance scores of 11-20. In the highest test score interval (41-50), 81 percent of the 166 individuals are in the top performance score interval, 18 percent in the next highest interval, and 1 percent in the middle interval.

In Figure 14, a performance score cutoff is applied, defining success, as before, as a performance score over 25. As with the expectancy chart, the expectancy table can provide an easy method of displaying the proportion of successes at various test intervals. Once again, both the different validities of samples A-C (Figures 14a-c) and the abnormal distributions in samples D and E (Figures 14d and e) are apparent.

A simple 2 x 2 expectancy table provides a quick indication of the proportion of successes when a cutoff test score is used. In Figures 15a-e, a test score cutoff of 25 is applied to each of the five sample data sets. For data set A (Figure 15a), 90 percent of the individuals who would be accepted would succeed, while for data set C (Figure 15c) only 60 percent of those accepted would succeed.

With the expectancy methods discussed above, the proportion of successes is evaluated. However, frequently it may be useful to look at the actual number of individuals at various test and performance score intervals. The frequency table is also a method of displaying information about the success or failure of individuals at various test score intervals. In the frequency table, both the test score and the performance measure are grouped into intervals and the numbers of individuals who fall into each interval are listed. The frequency table, in essence, provides a summary of the information in the scatter plot. Figures 16a-e show frequency tables using 10 five-point intervals for both test score and the performance measure for data sets A-E. As with other display methods, the relative validities of data sets A-C are apparent if Figures 16a-c are compared, and the abnormalities in data sets D and E are visible in Figures 16d and e.

In Figure 17, successful performance has been defined again as a performance score over 25, and the test score intervals have been reduced to five 10-point intervals.

The frequency table provides an excellent method for evaluating the cost

or utility of a test score cutoff. Once the cutoff has been determined, a simple 2 x 2 table can be used to display the number of successes and failures in the select and reject groups. With this method, it is simple to determine the number of correct decisions made about individuals. Accepting an individual who succeeds and rejecting an individual who would fail are correct decisions. These are called, respectively, true positives (*TP*) and true negatives (*TN*). Accepting an individual who fails is a false positive (*FP*) (the test incorrectly predicts success), and rejecting an individual who would succeed is a false negative (*FN*). The proportion of correct decisions is the total of true positives and true negatives divided by the total number of individuals (*n*):

$$(TP + TN)/(TP + TN + FP + FN)$$

or

$$(TP + TN)/n.$$

The general form of the 2 x 2 table is shown in Figure 18. True positives are entered in the upper right quadrant and true negatives in the lower left quadrant. Figures 19a-e show 2 x 2 tables for the sample data sets. The proportions of correct decisions range from .90 for sample A to .56 for C. The different types of errors for samples D and E are clear. For sample D (Figure 19d), there are virtually no false positives, but many false negatives. For sample E (Figure 19e), there are many false positives but no false negatives. Although the proportions of correct decisions are the same for the two distributions, the tests would have very different implications. If the cost of failure is high, data set D would be excellent for minimizing the number of individuals accepted who would fail. On the other hand, if the cost of failure is minimal but the payoff for success is high, the relationship in data set E would be preferable.

Display methods such as those described above are useful tools for describing test validity to the nonstatistician. In addition, however, these methods have a major advantage over the validity coefficient. The validity coefficient is extremely sensitive to changes in the range of test or performance scores. In many validity studies the individuals being measured have already been selected on some basis (often the test itself), so they do not represent the full range of abilities. It is very difficult to get performance measures on individuals at the lower levels because test users (employers, school administrators, etc.) are reluctant to admit every applicant just to evaluate test validity. As a result, it is only possible to estimate the validity coefficient from the preselected individuals available. This truncation of the range of test and performance scores artificially reduces the validity coefficient of the test.

For example, Figure 20 shows the results of preselection on data set A at

various test score cutoffs. Even at a cutting score of 20, the general trend in the distribution is present. However, Figure 21 shows the results of preselection on validity coefficients for the sample data sets by imposing test-score cutoffs at the 10, 20, 30, and 40 test-score levels. The effect of preselection is most dramatic for the moderate validity sample (B), in which the validity coefficient drops from .52 to .39 if the lowest cutoff score is applied. Although the coefficient of .52 for data set B would represent a respectable validity, if the range is further restricted by preselection, the validity coefficient plummets to .28 for a test-score cutoff of 20. Mathematical methods exist for correcting the validity coefficient for restriction of range and should be applied when this situation exists.

Examination of the scatter plot and box-and-whisker plots can indicate the presence of restriction of range due to preselection, as there will be very few individuals at the lower intervals. In addition, display methods are generally less dependent on range. Although restriction of range will be apparent, these methods look at performance as a function of test score intervals. As a result, the efficiency of a test is often detectable in these display methods when the validity coefficient is artificially reduced. For example, the trends in the relationships in all of the sample data sets are still apparent if the lowest two test-score intervals are deleted in all of the ten-interval test score figures presented here.

In summary, while the validity coefficient is an important part of test evaluation, alternative methods exist for displaying the test-criterion relationship. The scatter plot and the box-and-whisker plot are particularly useful in the identification of distribution abnormalities. In addition, the box-and-whisker plot provides an indication of the levels at which there is the most (or least) prediction error. Expectancy methods (chart, table, and plot) are essential in the evaluation of the prediction of two-level criteria (e.g., success versus failure), especially in terms of cost-benefit analysis. The frequency table also permits determination of the proportion of correct decisions.

Many of the methods presented here were originally developed for use in personnel testing. However, these methods are easily extended to any test-criterion situation. The final choice of display methods should depend on both test user needs and level of psychometric expertise.

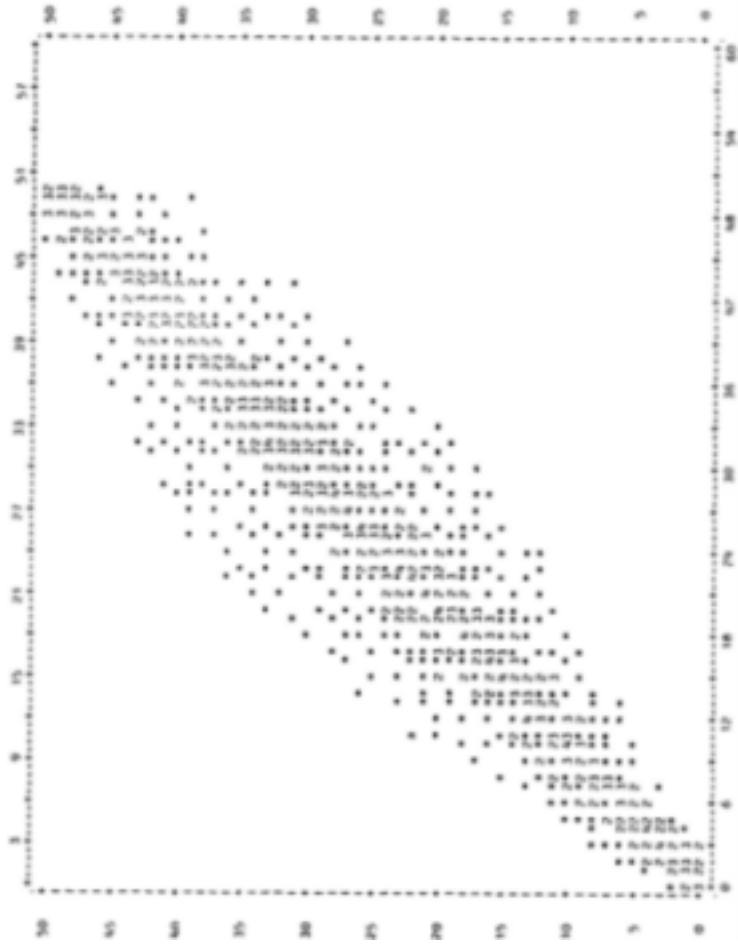


Figure 1a Scatter plot of performance score (down) by test score (across) for data set A.

Figure 1b Scatter plot of performance score (down) by test score (across) for data set B.

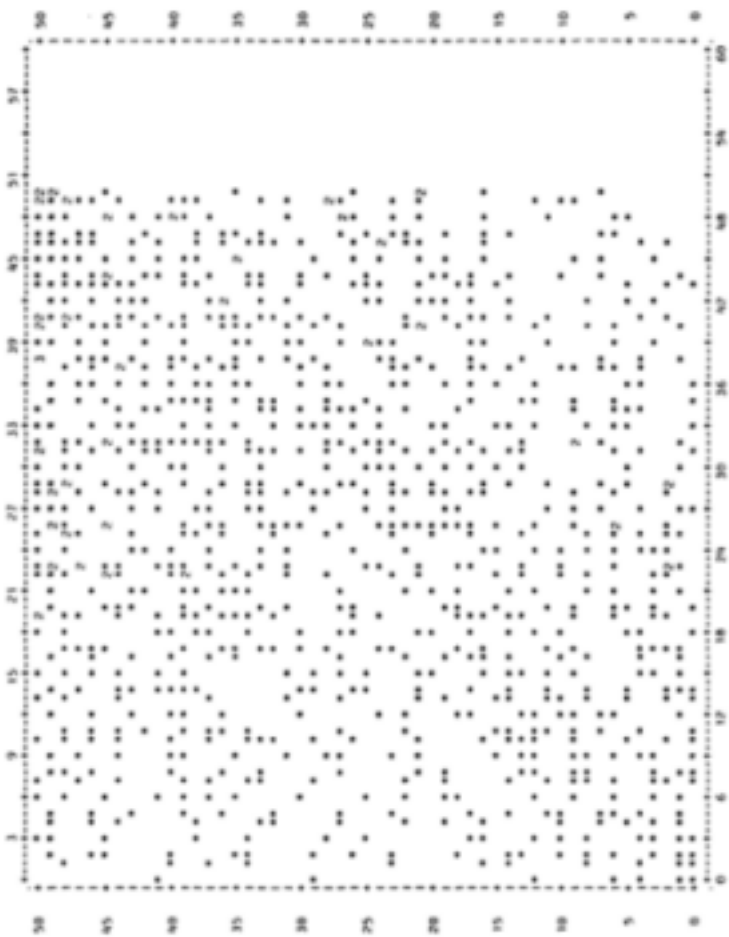


Figure 1c Scatter plot of performance score (down) by test score (across) for data set C.

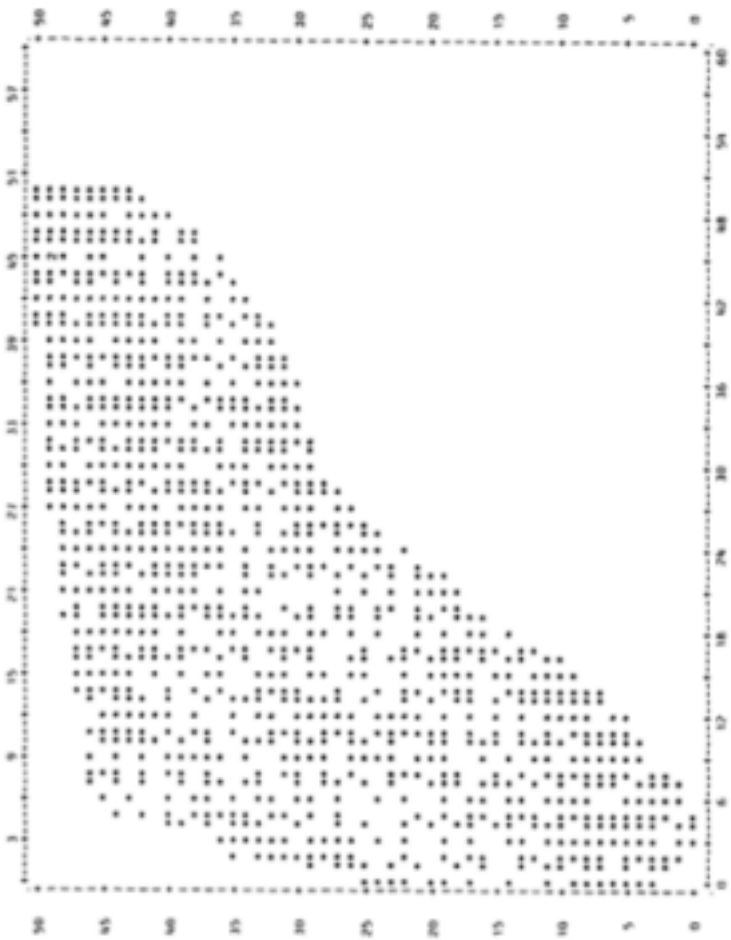


Figure 1d Scatter plot of performance score (down) by test score (across) for data set D.

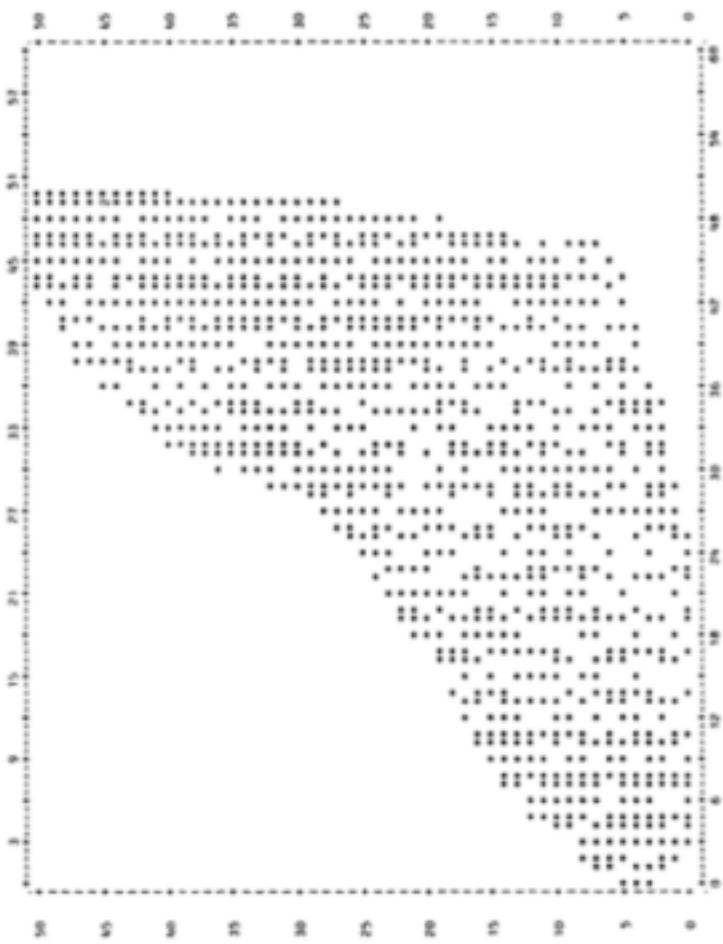
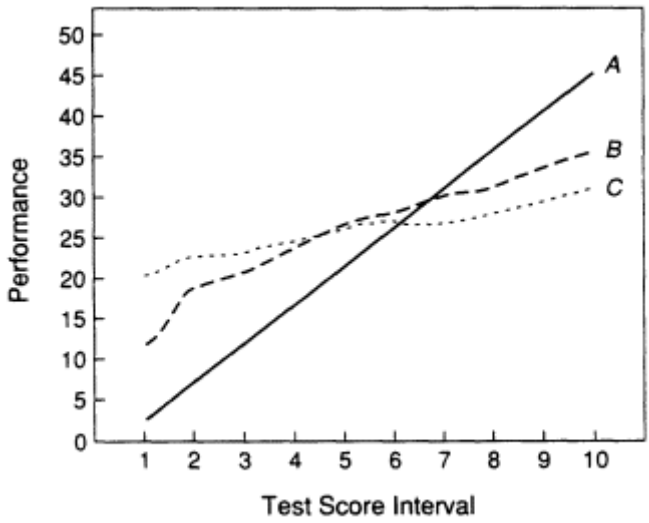


Figure 1e Scatter plot of performance score (down) by test score (across) for data set E.



Test Score Intervals	
1 = *0-5*	6 = *26-30*
2 = *6-10*	7 = *31-35*
3 = *11-15*	8 = *36-40*
4 = *16-20*	9 = *41-45*
5 = *21-25*	10 = *46-50*

Figure 2 Plot of mean performance score at five-point test score intervals for data sets A, B, C.

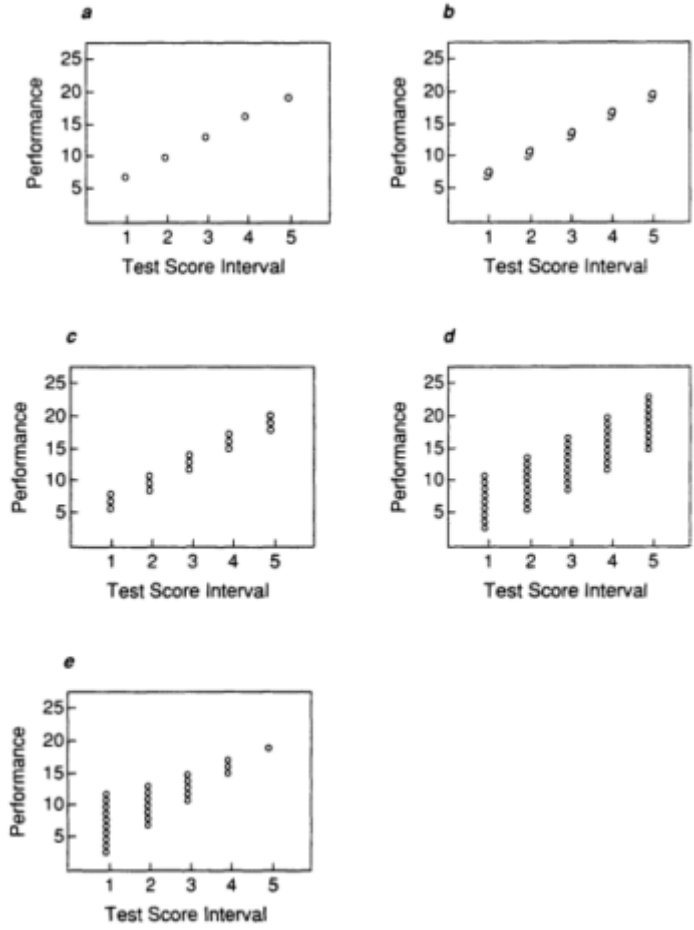


Figure 3 Plot of means with sample distributions.

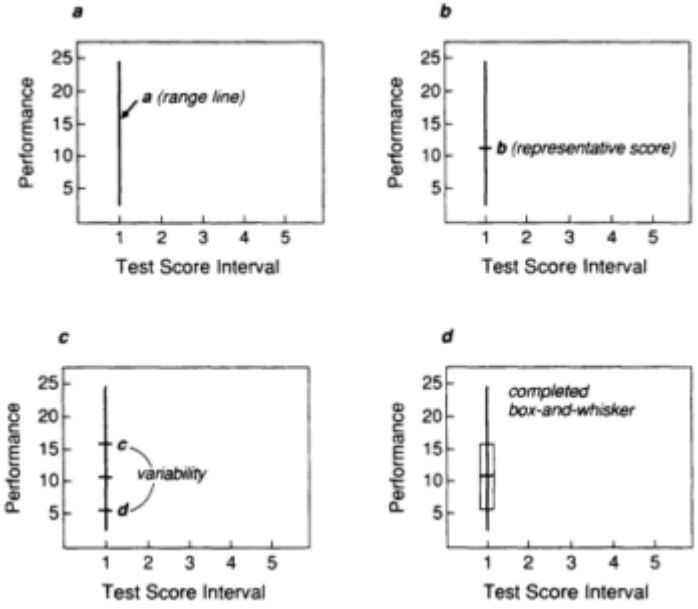


Figure 4 Construction of a box-and-whisker plot.

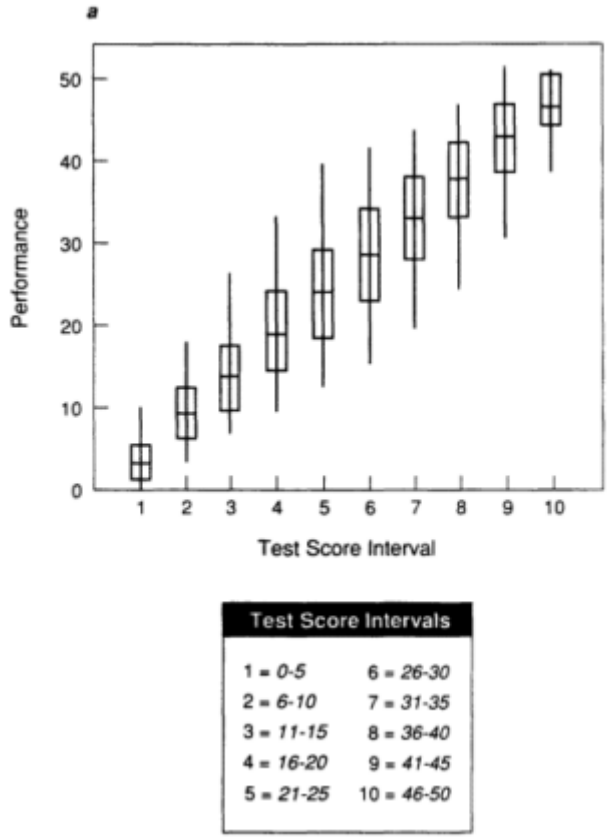


Figure 5a Box-and-whisker plot of data set A using the mean and standard deviation.

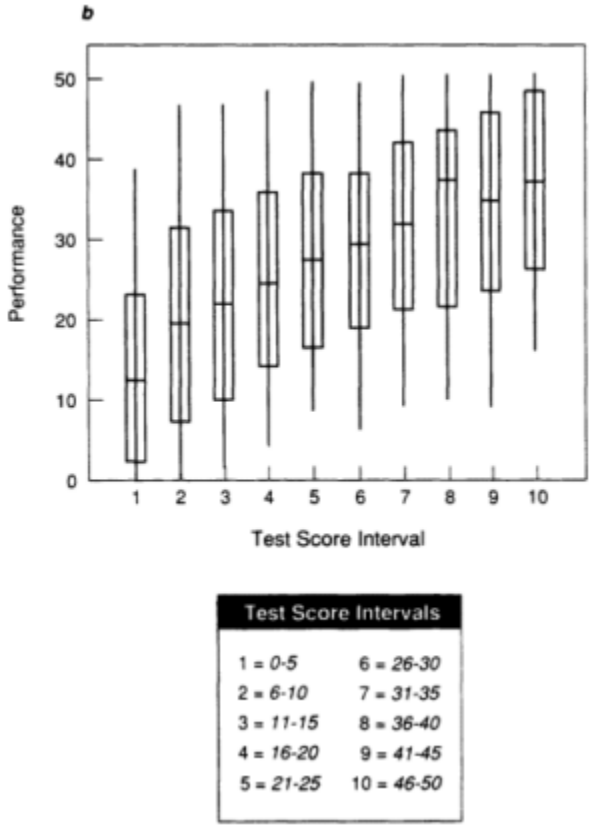


Figure 5b Box-and-whisker plot of data set B using the mean and standard deviation.

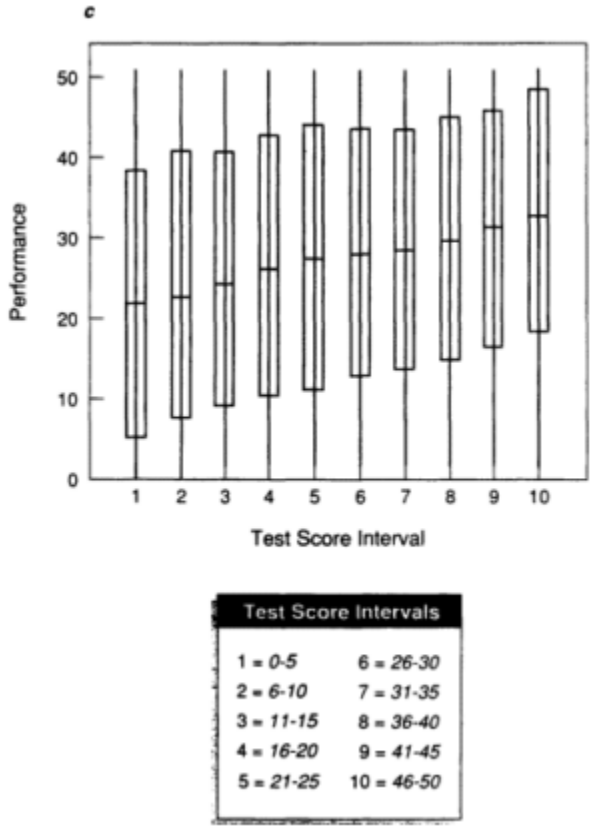


Figure 5c Box-and-whisker plot of data set C using the mean and standard deviation.

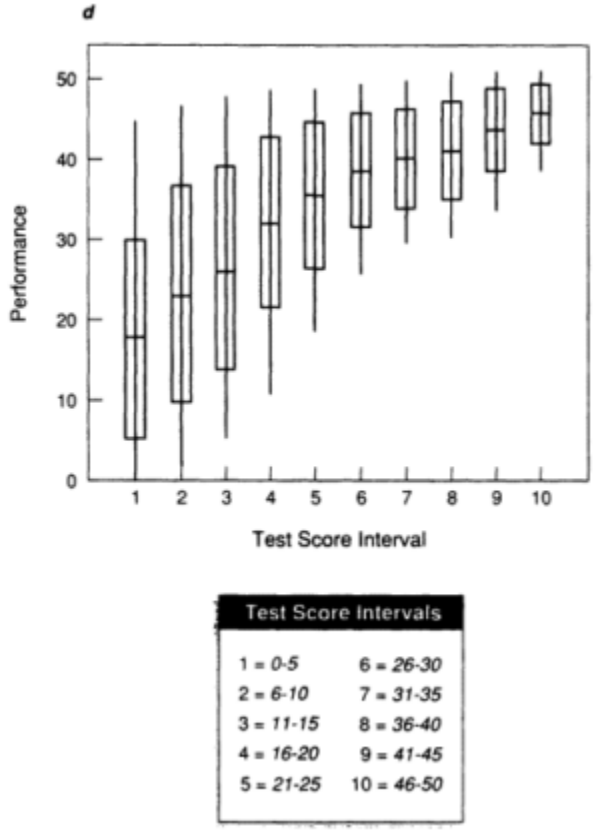


Figure 5d Box-and-whisker plot of data set D using the mean and standard deviation.

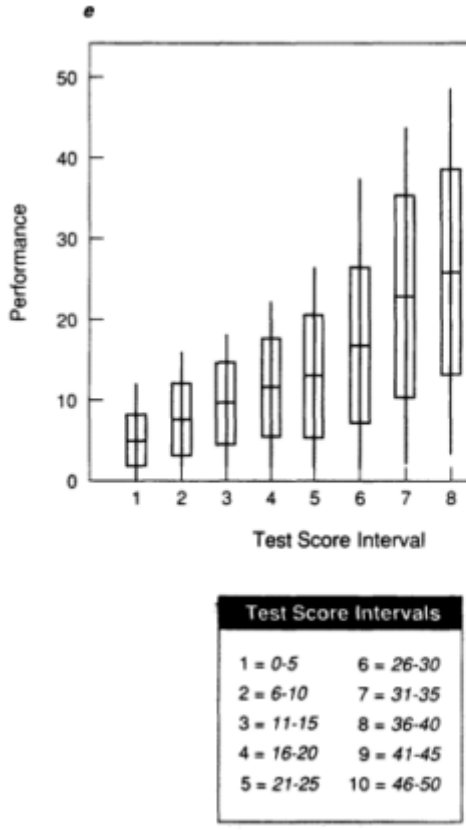
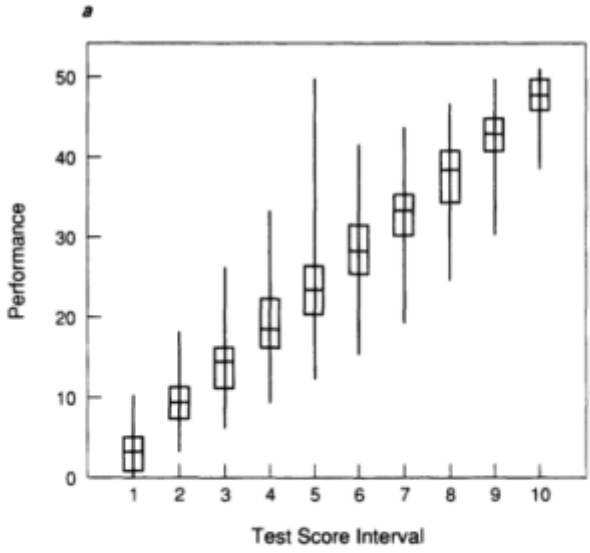


Figure 5e Box-and-whisker plot of data set E using the mean and standard deviation.



Test Score Intervals	
1 = *0-5*	6 = *26-30*
2 = *6-10*	7 = *31-35*
3 = *11-15*	8 = *36-40*
4 = *16-20*	9 = *41-45*
5 = *21-25*	10 = *46-50*

Figure 6a Box-and-whisker plot of data set A using the median and percentiles.

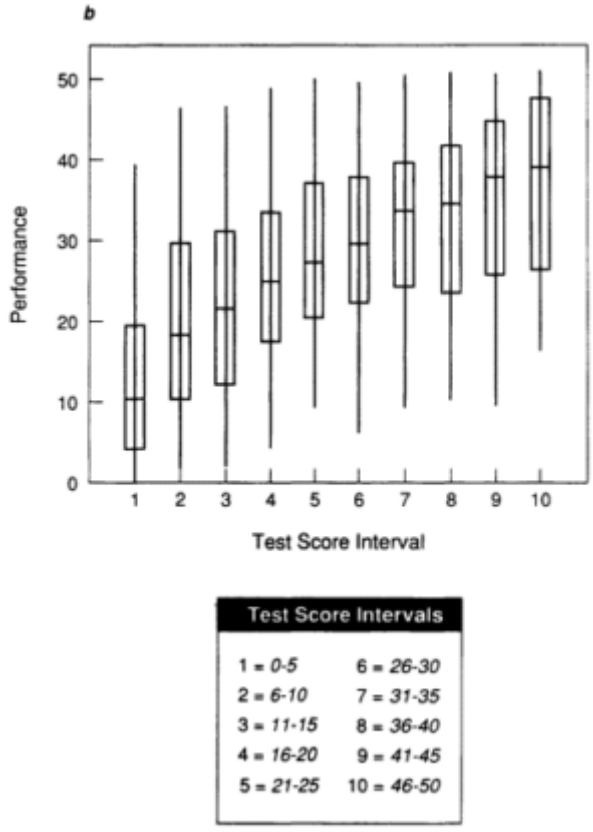


Figure 6b Box-and whisker plot of data set B using the median and percentiles.

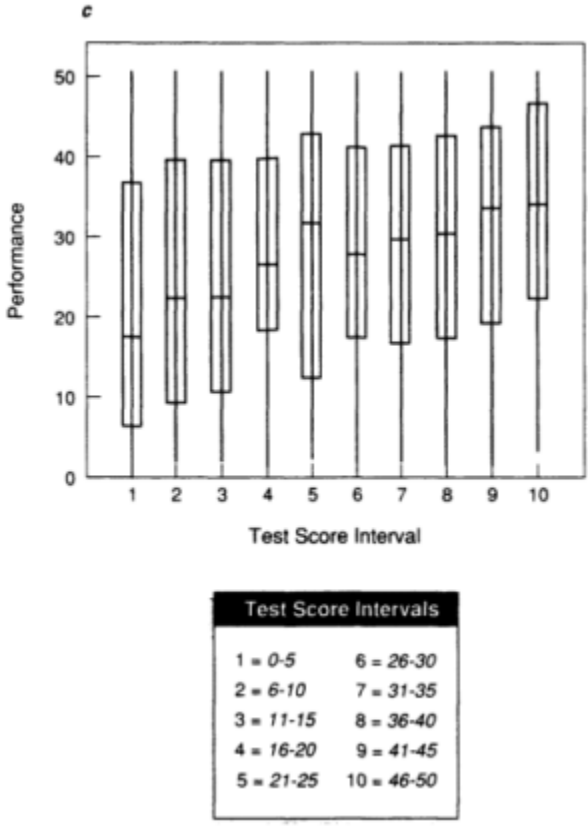


Figure 6c Box-and-whisker plot of data set C using the median and percentiles.

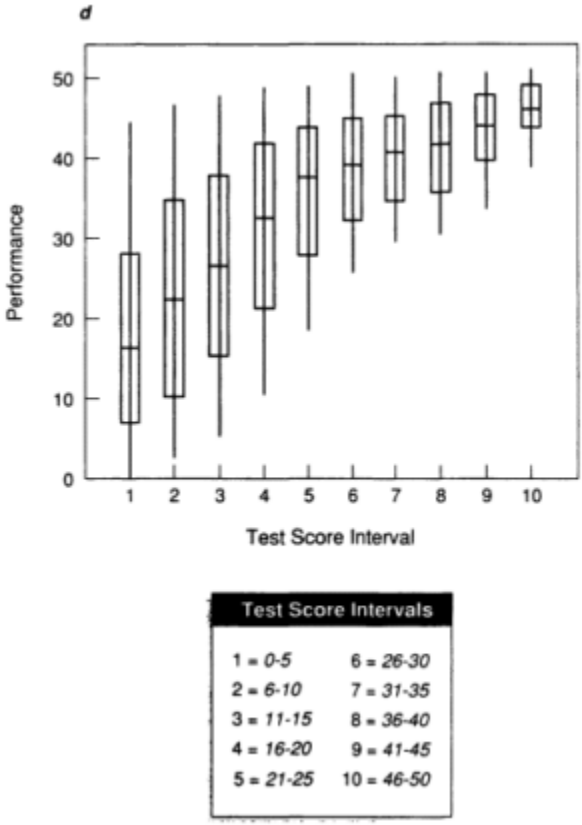
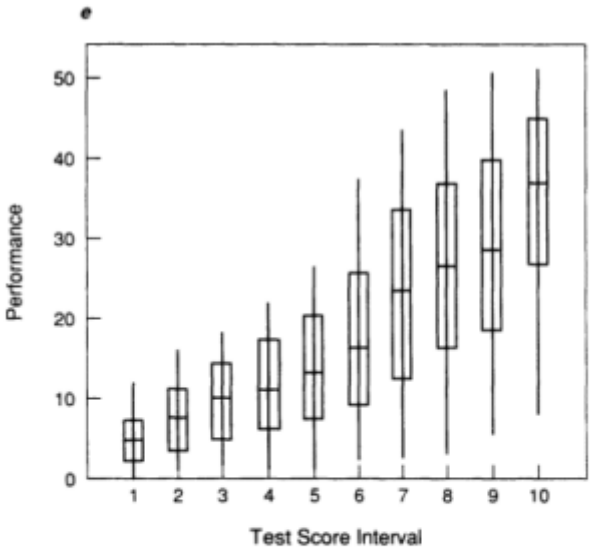


Figure 6d Box-and-whisker plot of data set D using the median and percentiles.



Test Score Intervals	
1 = *0-5*	6 = *26-30*
2 = *6-10*	7 = *31-35*
3 = *11-15*	8 = *36-40*
4 = *16-20*	9 = *41-45*
5 = *21-25*	10 = *46-50*

Figure 6e Box-and-whisker plot of data set E using the median and percentiles.

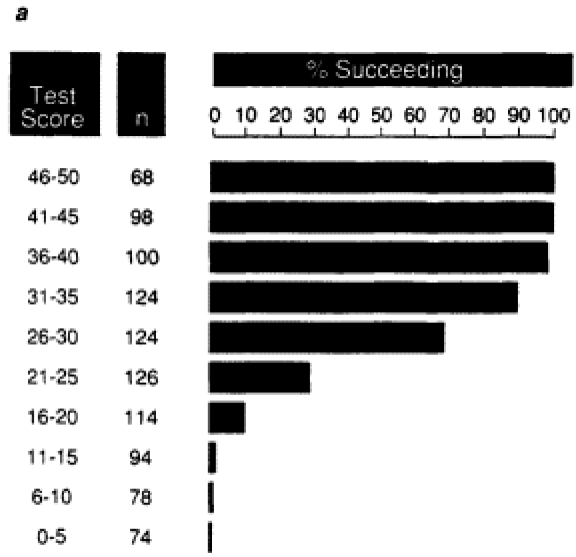


Figure 7a Expectancy chart for data set A: proportion of successful individuals in each test score interval.
Note: Success is defined as a performance score greater than 25.

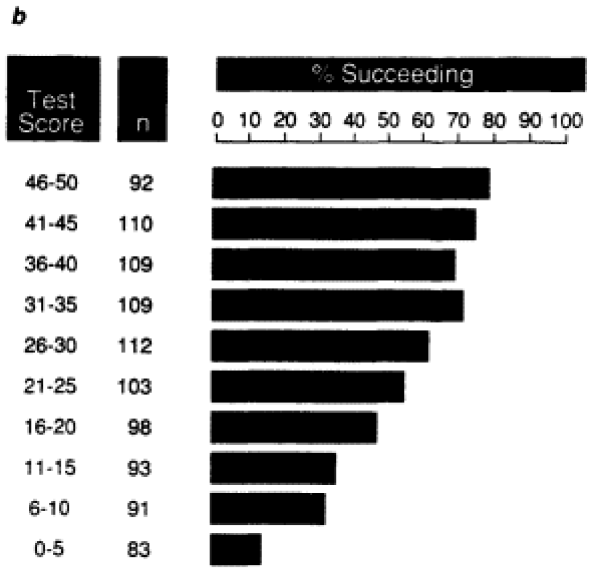


Figure 7b Expectancy chart for data set B: proportion of successful individuals in each test score interval.
Note: Success is defined as a performance score greater than 25.

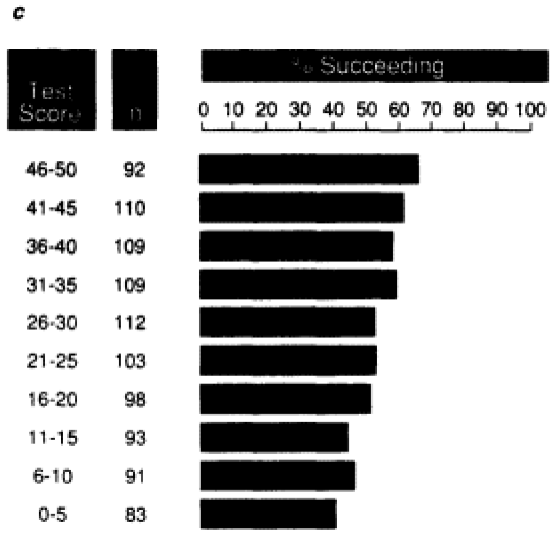


Figure 7c Expectancy chart for data set C: proportion of successful individuals in each test score interval.
Note: Success is defined as a performance score greater than 25.

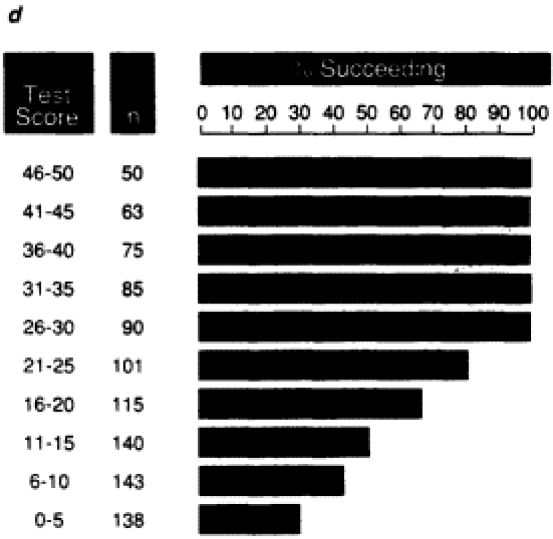


Figure 7d Expectancy chart for data set D: proportion of successful individuals in each test score interval.
Note: Success is defined as a performance score greater than 25.

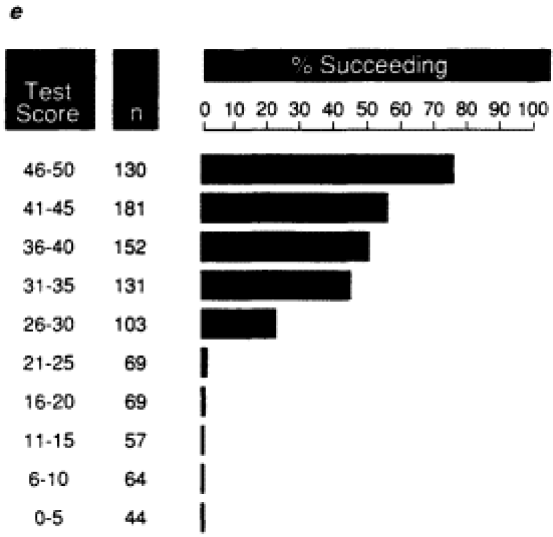


Figure 7e Expectancy chart for data set E: proportion of successful individuals in each test score interval.
Note: Success is defined as a performance score greater than 25.

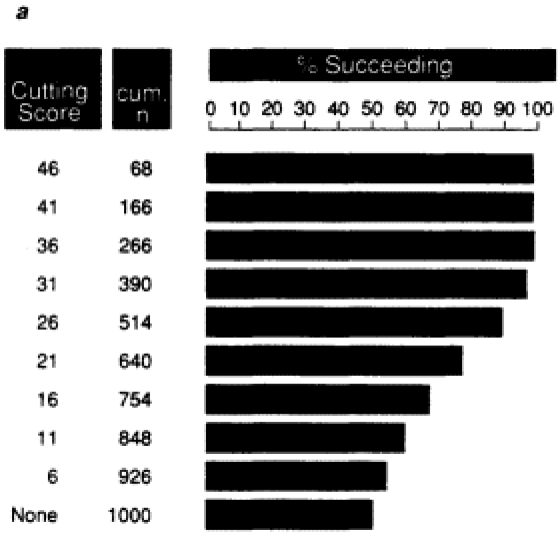


Figure 8a Expectancy chart for data set A: proportion of successful individuals at each cutting score.
Note: Success is defined as a performance score greater than 25.

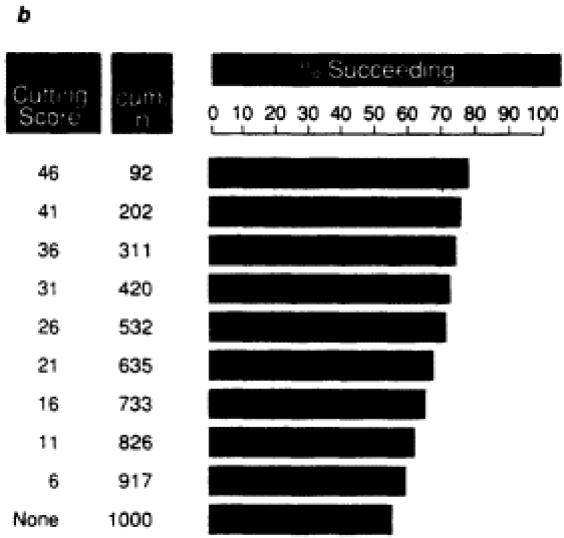


Figure 8b Expectancy chart for data set B: proportion of successful individuals at each cutting score.
Note: Success is defined as a performance score greater than 25.

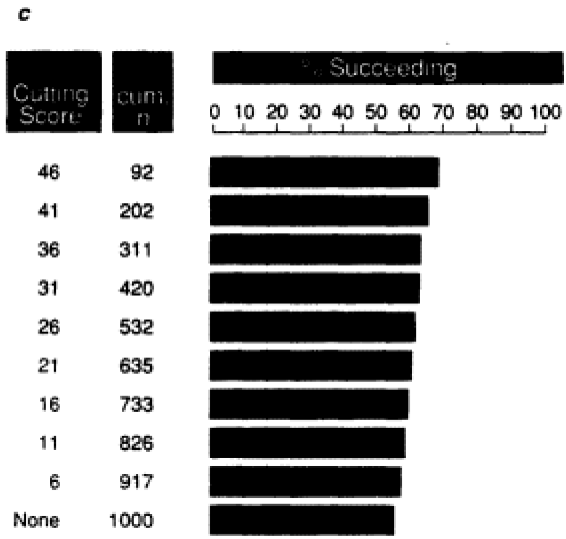


Figure 8c Expectancy chart for data set C: proportion of successful individuals at each cutting score.
Note: Success is defined as a performance score greater than 25.

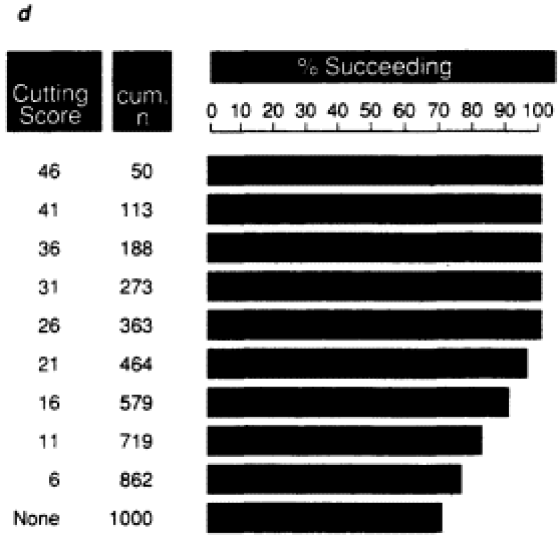


Figure 8d Expectancy chart for data set D: proportion of successful individuals at each cutting score.
Note: Success is defined as a performance score greater than 25.

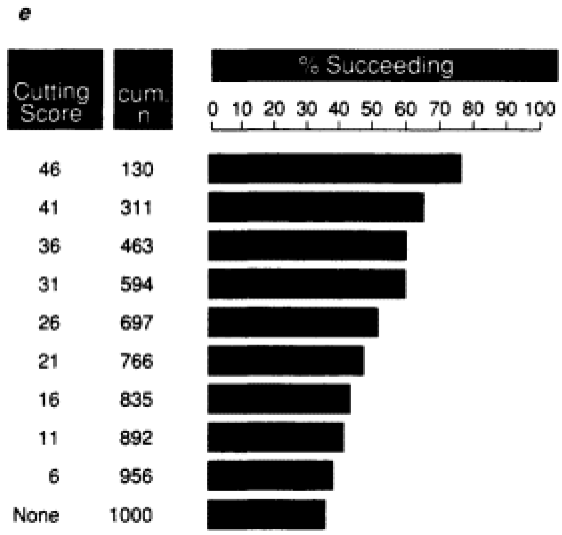


Figure 8e Expectancy chart for data set E: proportion of successful individuals at each cutting score.
Note: Success is defined as a performance score greater than 25.

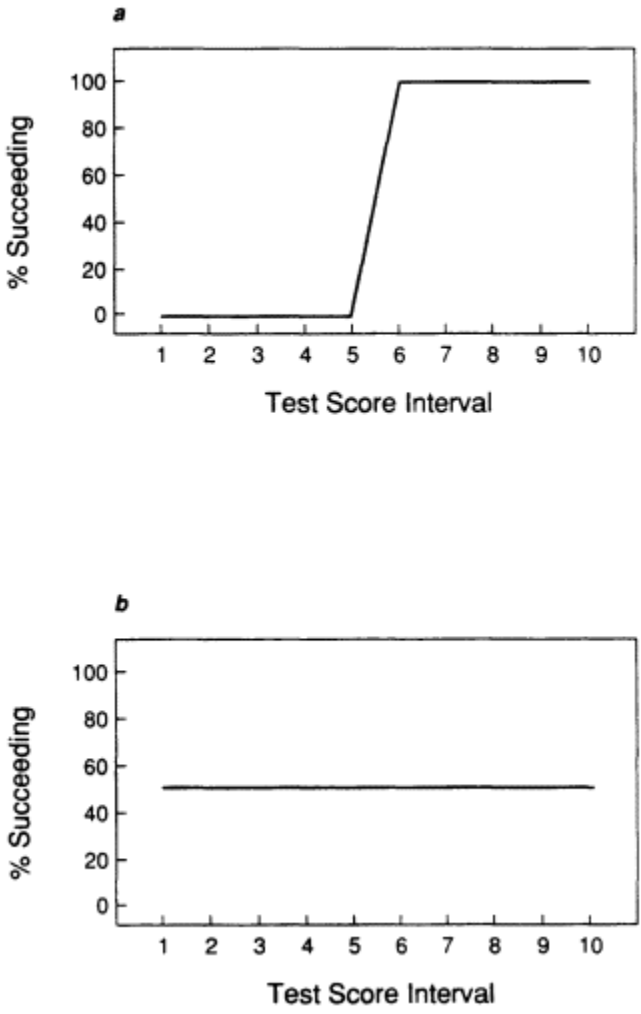


Figure 9 Plots of proportion of successful individuals in ten test score intervals for a perfectly discriminating test (9a) and a nondiscriminating test (9b).

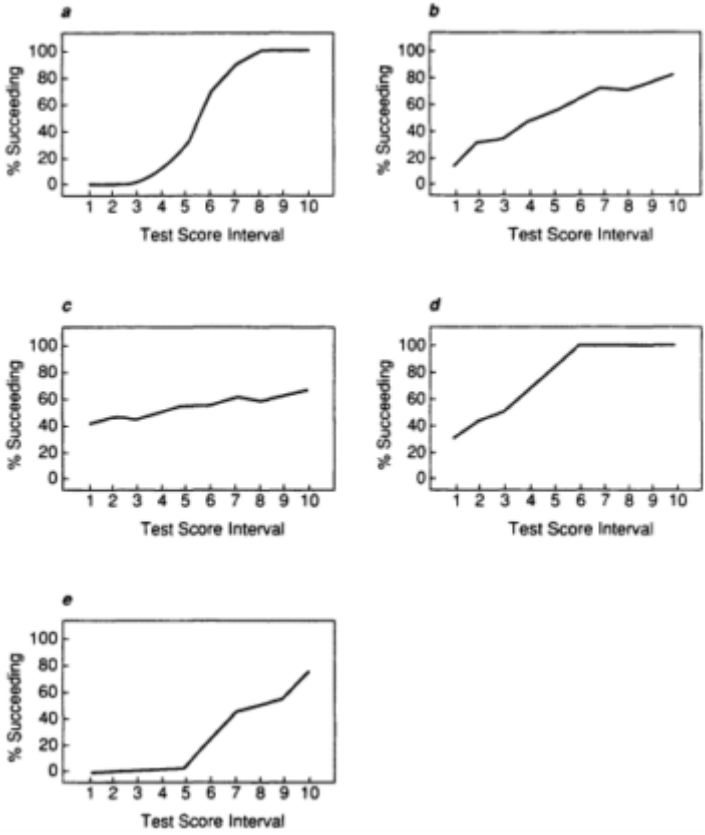


Figure 10 Plots of proportion of successful individuals in ten test score intervals for data sets A-E.

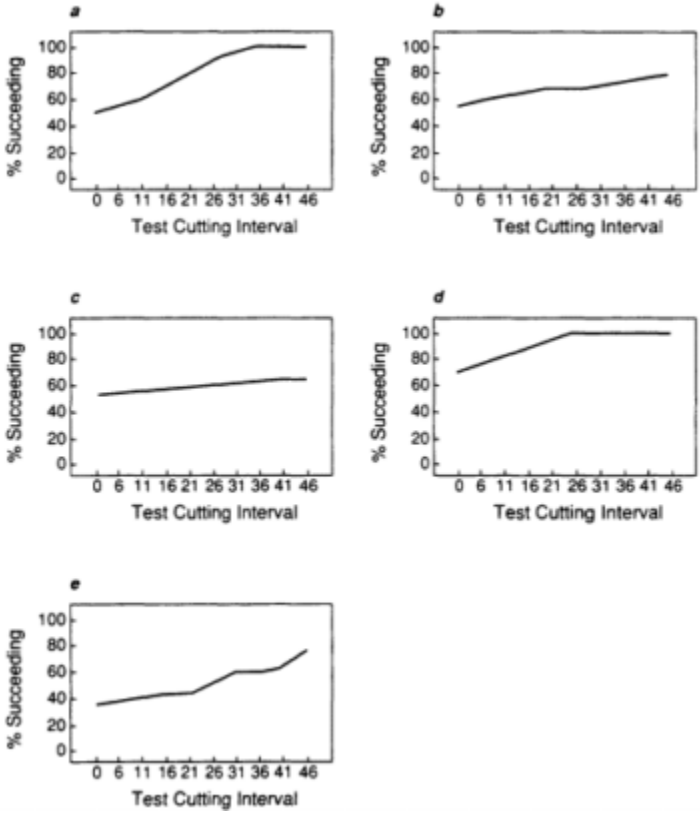


Figure 11 Plots of proportion of successful individuals at ten cutting scores on test for data sets A-E.

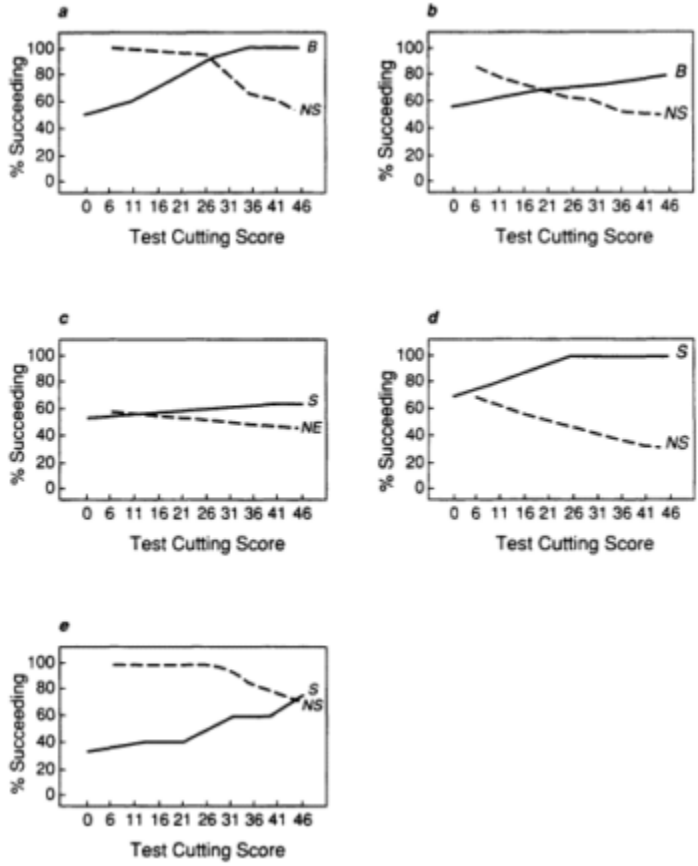


Figure 12 Proportions of individuals selected (S) who succeed and not selected (NS) who would have failed at ten cutting scores on test for data sets A-E.

a

Test Score	n	% at each criterion interval				
		0-10	11-20	21-30	31-40	41-50
41-50	166			1	18	81
31-40	224		1	21	65	13
21-30	250		19	62	18	.5
11-20	208	11	69	20	1	
0-10	152	87	13			

b

Test Score	n	% at each criterion interval				
		0-10	11-20	21-30	31-40	41-50
41-50	202	1	9	22	26	42
31-40	218	1	16	27	33	22
21-30	215	3	21	33	31	13
11-20	191	14	31	27	21	7
0-10	174	40	28	17	13	3

c

Test Score	n	% at each criterion interval				
		0-10	11-20	21-30	31-40	41-50
41-50	202	11	13	20	21	34
31-40	218	16	15	22	22	26
21-30	215	20	17	16	22	26
11-20	191	24	21	16	19	21
0-10	174	32	17	14	19	18

d

Test Score	n	% at each criterion interval				
		0-10	11-20	21-30	31-40	41-50
41-50	113				25	75
31-40	160			4	46	50
21-30	191		3	25	36	37
11-20	255	8	22	24	24	22
0-10	281	31	23	21	19	7

e

Test Score	n	% at each criterion interval				
		0-10	11-20	21-30	31-40	41-50
41-50	311	6	18	23	26	27
31-40	283	18	21	24	26	10
21-30	172	34	35	27	3	
11-20	126	51	45	4		
0-10	108	81	19			

Figure 13 Expectancy tables showing percent receiving each criterion score for five intervals of test and criterion.

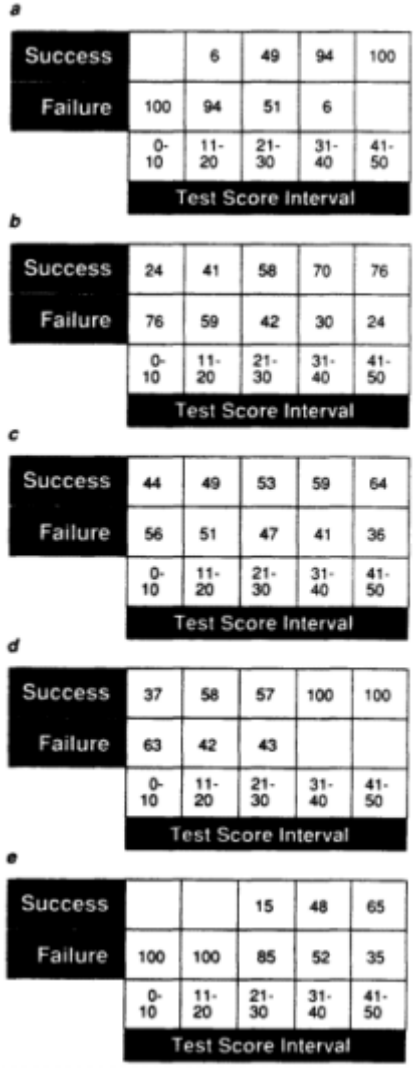


Figure 14 Expectancy tables showing percent succeeding and failing at five test score intervals for data sets A-E.
Note: Success is defined as a performance score greater than 25.

Figure 15 Expectancy tables showing percent succeeding and failing at test cutting score of 25 for data sets A-E.
Note: Success is defined as a performance score greater than 25.

a

Performance										
10								2	16	47
9						1	6	21	54	18
8					4	10	21	46	20	3
7				2	10	22	58	21	7	
6			1	9	23	53	27	9	1	
5			7	24	54	26	9	1		
4		3	19	54	27	11	3			
3		17	48	22	8	1				
2	14	49	19	3						
1	60	9								
	1	2	3	4	5	6	7	8	9	10
Test Score Interval										

Test and Performance Intervals	
1 = *0-5*	6 = *26-30*
2 = *6-10*	7 = *31-35*
3 = *11-15*	8 = *36-40*
4 = *16-20*	9 = *41-45*
5 = *21-25*	10 = *46-50*

Figure 16a Frequency tables for ten intervals of test score and performance score for data set A.

b

Performance \ Test Score Interval	1	2	3	4	5	6	7	8	9	10
10		1	2	3	5	6	9	13	22	27
9		4	5	4	6	10	11	16	22	13
8	3	9	11	10	16	18	23	17	15	13
7	4	6	5	14	15	17	27	16	17	8
6	5	9	10	15	14	18	18	13	6	11
5	6	9	14	13	20	19	12	15	11	17
4	10	11	13	19	12	14	8	12	11	3
3	11	17	17	10	11	8	9	6	4	
2	21	15	11	7	4	2	2	1	2	
1	23	10	5	3						

Figure 16b Frequency tables for ten intervals of test score and performance score for data set B.

c

Performance \ Test Score Interval	1	2	3	4	5	6	7	8	9	10
10	9	12	11	12	16	18	13	17	21	23
9	5	5	9	8	11	10	14	12	16	9
8	7	10	11	12	15	11	13	13	12	11
7	7	9	3	10	10	11	11	12	11	9
6	6	7	8	9	4	9	14	10	8	10
5	5	7	8	5	8	14	9	14	10	13
4	7	5	6	12	7	14	9	9	13	5
3	7	11	14	8	8	7	8	6	5	3
2	12	11	10	8	11	8	10	8	5	6
1	18	14	13	14	13	10	8	8	9	3

Figure 16c Frequency tables for ten intervals of test score and performance score for data set C.

d

Performance											
10		3	6	9	13	16	16	19	23	25
9	2	14	19	23	21	20	24	21	17	20
8	8	16	13	16	21	21	18	16	18	5
7	16	12	16	15	13	14	22	18	5	
6	16	17	18	14	14	18	5	1		
5	13	13	15	13	14	1				
4	16	16	15	17	5					
3	16	16	18	7						
2	25	20	18	1						
1	26	16	2							
	1	2	3	4	5	6	7	8	9	10
Test Score Interval										

Figure 16d Frequency tables for ten intervals of test score and performance score for data set D.

e

Performance											
10								5	20	25
9							5	18	17	23
8						1	18	16	21	19
7						5	22	18	24	16
6					1	18	14	21	20	16
5				5	14	14	13	21	21	13
4		1	7	17	13	14	15	16	23	9
3	2	18	18	15	15	18	16	13	19	6
2	16	20	16	16	13	17	12	16	14	3
1	26	25	16	16	13	16	16	8	2	
	1	2	3	4	5	6	7	8	9	10
Test Score Interval										

Figure 16e Frequency tables for ten intervals of test score and performance score for data set E.

a

Success		12	123	211	166
Failure	152	196	127	13	
	0-10	11-20	21-30	31-40	41-50
	Test Score Interval				

b

Success	41	79	125	153	154
Failure	133	112	90	65	48
	0-10	11-20	21-30	31-40	41-50
	Test Score Interval				

c

Success	77	93	115	129	130
Failure	97	98	100	89	72
	0-10	11-20	21-30	31-40	41-50
	Test Score Interval				

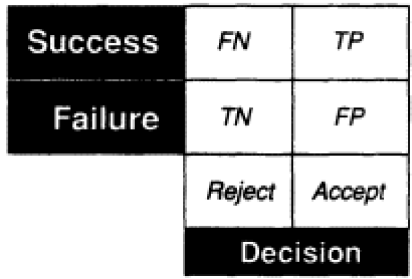
d

Success	104	149	108	160	113
Failure	177	106	83		
	0-10	11-20	21-30	31-40	41-50
	Test Score Interval				

e

Success			25	137	201
Failure	108	126	147	146	110
	0-10	11-20	21-30	31-40	41-50
	Test Score Interval				

Figure 17 Frequency tables showing number succeeding and failing at five test score intervals for data sets A-E.
Note: Success is defined as a performance score greater than 25.



TP = true positive *FP = false positive*
TN = true negative *FN = false negative*

$$\text{Proportion of Correct Decisions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Figure 18 General form of the frequency table, computation of proportion of correct decisions is shown.

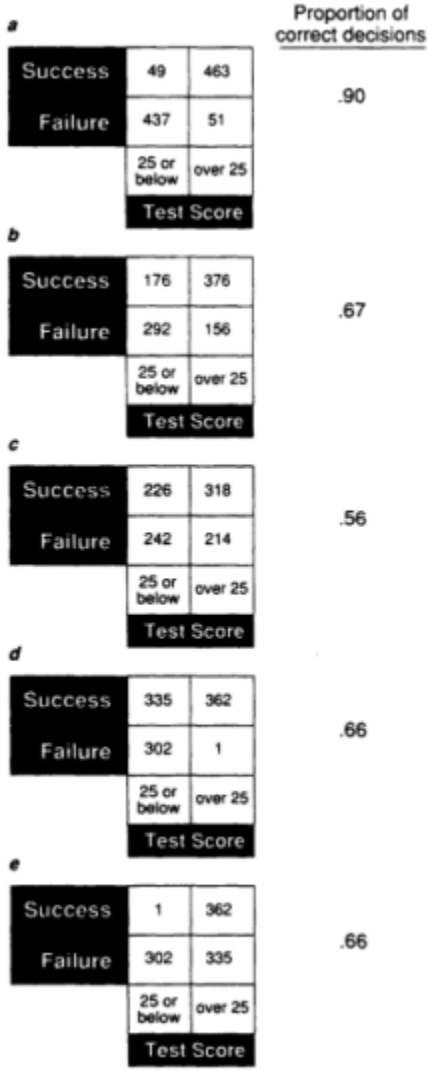


Figure 19 Frequency tables showing number succeeding and failing at test cutoff score for 25 for data ses A-E.
Note: Success is defined as a performance score greater than 25.

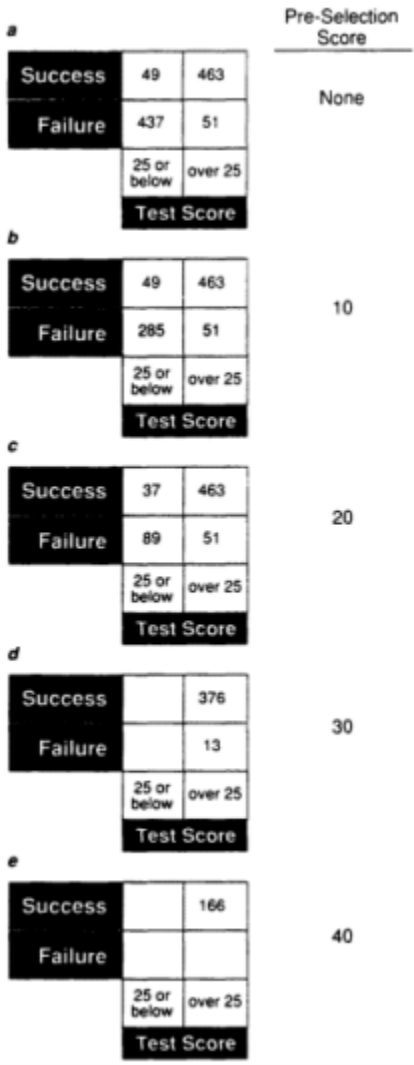


Figure 20 Frequency tables showing effect of preselection of frequency table in Figure 19 for data set A.
Notes: Success is defined as a performance score greater than 25. Test cutoff score of 25 determines levels of test score.

	Pre-test Selection Cutoff				
Data Set	*None*	*10*	*20*	*30*	*40*
A	.95	.91	.86	.80	.62
B	.52	.39	.28	.20	.13
C	.19	.15	.11	.11	.07
D	.65	.56	.44	.39	.31
E	.65	.58	.49	.33	.28

Figure 21 Effect of restriction of range by preselection on validity coefficients for data sets A-E.

Generalizability Theory and Military Performance Measurements: I. Individual Performance

Richard J. Shavelson

INTRODUCTION

This paper sketches a statistical theory of the multifaceted sources of error in a behavioral measurement. The theory, generalizability (G) theory (Cronbach et al., 1972), models traditional measurements such as aptitude and achievement tests. It provides estimates of the stability of a measurement ("test-retest" reliability in classical test theory), the consistency of responses to parallel forms of a test ("equivalent-forms" reliability), and the consistency of responses to test items ("internal-consistency" reliability). Each type of classical reliability coefficient defines measurement error somewhat differently. One of G theory's major achievements is that it simultaneously estimates the magnitude of the errors influencing all three classical reliabilities. Hence, we speak of G theory as a theory of the multifaceted sources of error.

Performance measurements may contain the same sources of error as traditional pencil-and-paper measurements: instability of responses from one occasion to the next, nonequivalence of supposedly parallel forms of a performance measurement, and heterogeneous subtask responses. And more. Two additional, pernicious sources of error are inaccuracies due to scoring, where observers typically score performance in real time, and inaccuracies

The author gratefully acknowledges helpful and provocative comments provided by Lee Cronbach and the graduate students in his seminar on generalizability theory. The author alone is responsible for the contents of this paper.

due to unstandardized testing conditions, where performance testing is typically carried out under widely varying laboratory and field conditions.[1] G theory's ability to estimate the magnitude of each of these sources of error, individually and in combinations, enables this theory to model human performance measurement better than any other.

The next section provides an example of how generalizability theory can be applied to military job performance measurements, using hypothetical data. The third section presents G theory formally, but with a minimum of technical detail. Key features of the theory are illustrated with concrete numerical examples. The fourth section presents applications of the theory. These applications were chosen to highlight the theory's flexibility in modeling a wide range of measurements. The fifth section concludes the paper by discussing some limitations of the theory.

APPLICATION OF GENERALIZABILITY THEORY TO THE MEASUREMENT OF MILITARY PERFORMANCE

Background

Military decision makers, ideally, seek perfectly reliable measures of individuals' performance in their military occupational specialities.[2] Even with imperfect measures, the decision maker typically treats as interchangeable measures of an individual's performance on one or another representative sample of military occupational specialty tasks (and subtasks) that were carried out at any one of many test stations, on any of a wide range of occasions, as scored by any of a large number of observers. Because he wants to know what the person's performance is like, rather than what he did on one particular moment of observation, he is forced to *generalize* from a limited sample of behavior to an extremely large universe: the individual's job performance across time, tasks, observers, and settings. This inference is sizable. Generalizability theory provides the statistical apparatus for answering the question: Just how dependable is this measurement-based inference?

To estimate dependability, an individual's performance needs to be observed on a sample of tasks/subtasks, on different occasions, at different stations, with different observers. A generalizability study (G study), then, might randomly sample five E-2s,[3] who would perform a set of tasks (and subtasks) on two different occasions, at two different stations, with four

[1] By design, traditional pencil-and-paper tests control for scoring errors by using a multiple-choice format with one correct answer, and testing conditions are standardized by controlling day, time of day, instructions, etc.

[2] "Military occupational specialty" is used generically and applies to Air Force specialties and Navy ratings as well as to Army and Marine Corps military occupational specialties.

[3] Large samples should be used. For illustrative purposes, small samples are more instructive.

observers scoring their performance. An individual would be observed under all possible combinations of these conditions or a total of 16 times (2 occasions x 2 stations x 4 observers) on the set of tasks/subtasks.

If performance is consistent across tasks, occasions, stations, and observers—i.e., if these characteristics of the measurement do not introduce systematic or unsystematic variation in the measurement—the measurement is dependable and the decision maker's ideal has been met. More realistically, however, if the individual's score depends on the particular sample of tasks to which he was assigned, on the particular occasion or station at which the measurement was taken, and/or on the particular observer scoring the performance, the measurement is less than ideally dependable. In this case, interest attaches to determining how to minimize the impact of different sources of measurement error.

Performance Measurement: Operate and Maintain Caliber .38 Revolver

To make this general discussion concrete, an example is in order. One of the Army's military occupational specialty-specific performance measures involves operating and maintaining a caliber .38 revolver. The soldier is told that this task covers the ability to load, reduce a stoppage in, unload, and clean the caliber .38 revolver, and that this will be timed. The score sheet for this measurement is presented in Table 1. Note that there are two measurements taken: time and accuracy.

In the G study, suppose that each of five soldiers performed the revolver test four times: on two different occasions (e.g., week 1 and week 2) at two different test stations.[4] The soldiers' performance on each of the three tasks and subtasks (see Table 1) was independently scored by four observers. Also, each task as a whole is independently timed. Hypothetical results of this study are presented in Table 2 for the time measure. Note that time is recorded for each of three tasks and not for individual subtasks (Table 1); hence, subtasks are not shown in Table 2.

Classical Theory Approach

With all the information provided in Table 2, how might classical reliability be calculated? With identical performance measurements taken on

[4] There is good reason to worry about an order effect. This is why "tuning" subjects before they are tested is strongly recommended (e.g., Shavelson, 1985). "Tuning" is familiarizing subjects with the task before they are tested. (If a subject can "fake" the task in a performance test, this means that she can perform it.) Nevertheless, soldiers would be counterbalanced such that half would start at station I and half at station 2. Finally, as will be seen, an alternative design with occasions nested within stations might be used.

TABLE 1 Caliber .38 Revolver Operation and Maintenance Task

		Score	
Task	Subtask	Go	No Go
Load the weapon[a]	(1) Held the revolver forward and down	-	-
	(2) Pressed thumb latch and pushed cylinder out	-	-
	(3) Inserted a cartridge into each chamber of the cylinder	-	-
	(4) Closed the cylinder	-	-
	(5) Performed steps 1-4 in sequence	-	-
	Time to load the weapon_____		
Reduce a stoppage[b]	(6) Recocked weapon	-	-
	(7) Attempted to fire weapon	-	-
	(8) Performed steps 6-7 in sequence	-	-
	Time to reduce stoppage_____		
Unload and clear the weapon[c]	(9) Held the revolver with muzzle pointed down	-	-
	(10) Pressed thumb latch and pushed cylinder out	-	-
	(11) Ejected cartridges	-	-
	(12) Inspected cylinder to ensure each chamber is clear	-	-
	(13) Performed steps 6-9 in sequence	-	-
	Time to unload and clear the weapon_____		

NOTES: Instructions to soldier:

[a] This task covers your ability to load the revolver; we will time you. Begin loading the weapon.

[b] You must now apply immediate action to reduce a stoppage. Assume that the revolver fails to fire. The hammer is cocked. Begin.

[c] You must now begin unloading the weapon.

TABLE 2 Caliber .38 Revolver Operation and Maintenance Task: Time to Complete Tasks

			Observer			
Station	Occasion	Task	1	2	3	4
1			84	85	86	87
			82	84	85	85
		1	91	92	92	94
			83	82	84	85
			75	76	78	78
			76	76	77	77
			75	84	75	76
	1	2	83	81	83	81
			77	78	76	77
			69	70	70	70
			94	95	96	97
			91	92	93	94
		3	99	99	99	99
			93	94	94	95
			83	83	84	85
				* * *		
2			80	81	81	82
			78	78	81	80
		1	84	84	84	85
			80	81	80	82
			73	74	74	75
			73	73	74	76
			74	73	74	75
	2	2	77	75	76	75
			73	74	72	77
			69	70	70	71
			90	89	90	92
			90	89	90	91
		3	89	91	93	93
			87	87	89	89
			83	84	85	84

two occasions, a test-retest reliability can be calculated. By recognizing that tasks are analogous to items on traditional tests, an internal consistency reliability coefficient can be calculated.

A test-retest coefficient is calculated by correlating the soldiers' scores at occasion 1 and occasion 2, after summing over all other information in Table 2. The correlation between scores at the two points in time is .97. If soldiers' performance times are averaged over two occasions to provide a performance time measure, the reliability is .99, following the Spearman-Brown prophecy formula.

An internal-consistency coefficient is calculated by averaging, for each task, soldiers' performance times across stations, occasions, and observers. The soldiers' average task performance times would then be intercorrelated: r(task 1, task 2), r(task 1, task 3), and r(task 2, task 3). The average of the three correlations would provide the reliability for a single task, and the Spearman-Brown formula could be used to determine the reliability for performance times averaged over the three tasks. The reliability of performance-time measures obtained on a single task is .99, and the reliability of scores averaged across the three tasks is .99.

Generalizability Theory Approach

Two limitations of classical theory are readily apparent. The first limitation is that a lot of information in Table 2 is ignored (i.e., "averaged over"). This information might contain measurement error that classical theory assumes away. This could lead to false confidence in the dependability of the performance measure. The second limitation is that separate reliabilities are provided; which is the "right one"? G theory overcomes both limitations. The theory uses all of the information obtained in the G study, and it provides a coefficient that includes a definition of error arising from each of the sources of error in the measurement. Finally, G theory estimates each source of variation in the measurement separately so that improvements can be made by pinpointing which characteristics of the performance measurement gave rise to the greatest error.

Generalizability theory uses the analysis of variance (ANOVA) to accomplish this task. A measurement study (called a generalizability study) is designed to sample potential sources of measurement error (e.g., raters, occasions, tasks) so that their effects on soldiers' performance can be examined. Thus soldiers and each source of error can be considered factors in an ANOVA. The ANOVA, then, can be used to estimate the effects of soldiers (systematic, "true-score" variation), each source of error, and their interactions. More specifically, the ANOVA is used to estimate the variance components associated with each effect in the design ("main effects" and "interactions"). As Rubin (1974:1050) noted, G theory concentrates on mixed models analysis

of variance designs, that is, designs in which factors are crossed or nested and fixed or random. Emphasis is given to the estimation of variance components and ratios of variance components, rather than the estimation and testing of effects for fixed factors as would be appropriate for designs based on randomized experiments.

Variance Components

The statistical machinery for analyzing the results of a G study is the analysis of variance. The ANOVA partitions the multiple sources of variation into separate components ("factors" in ANOVA terminology) corresponding to their individual main effects (soldiers, stations, occasions, tasks, and judges) and their combinations or interactions. The total variation in performance times (shown in Table 2) is partitioned into no less than 31 separate components—five individual components and all their possible combinations (Cartesian products)—accounting for the total variation in the performance-time data (see Table 3).

Of the 30 sources of variation, 1 accounts for performance consistency: the *soldier* (or P for person) effect represents systematic differences in the speed of performance among the five soldiers (variance component for soldiers in Table 3). By averaging the time measure across observers, tasks, occasions, and stations, we find that soldier 5 performed the task the fastest and soldier 3 performed the task the slowest. The other three soldiers fell in between. This variation in mean performance can be used to determine systematic differences among soldiers, called true-score variance in classical test theory and universe-score variance in generalizability theory. This universe-score variance—variance component for P = 14.10 (Table 3)—is the signal sought through the noise created by error. It is the "stuff' that the military decision maker would like to know as inexpensively and as feasibly as possible.

The 29 other sources of variation represent potential measurement error. The first four sources of variation are attributable to each source of error considered singly ("main effects" in ANOVA terminology). The *station* effect (variance component for station in Table 3) shows whether mean performance times, averaged over all other factors, systematically vary as to the location at which the measurement was taken. Apparently performance time did not differ according to station (variance component for station = 0). This is not surprising; unlike many other performance measurements, the revolver task appears self-contained. The *occasion* effect shows whether performance times, averaged over all other factors, change from one occasion to the next. Relative to other variance components, performance appears stable over occasions. The *task* effect shows whether performance times differed over tasks 1-3. Since task 2 contained fewer subtasks (three)

TABLE 3 Generalizability Study for a Soldier (P) x Station (S) x Occasion (0) x Task (T) x Judge (J) Design

Source of Variation	df	Mean Squares	Variance Components
Soldiers (*P*)	4	1020.80	14.10
Stations (*S*)	1	1.00	0.00
Occasions (*O*)	1	1273.00	7.40
Tasks (*T*)	2	1659.80	20.00
Judges (*J*)	3	349.80	2.45
PS	4	1.00	0.00
PO	4	239.00	9.55
PT	8	9.80	0.00
PJ	12	106.80	8.75
SO	1	1.00	0.00
ST	2	1.00	0.00
SJ	3	1.00	0.00
OT	2	59.80	1.25
OJ	3	97.80	3.20
TJ	6	1.80	0.00
PSO	4	1.00	0.00
PST	8	1.00	0.00
PSJ	12	1.00	0.00
POT	8	9.80	1.00
POJ	12	1.80	0.00
PTJ	24	1.80	0.00
SOT	2	1.00	0.00
SOJ	3	1.00	0.00
STJ	6	1.00	0.00
OTJ	6	1.80	0.00
PSOT	8	1.00	0.00
PSOJ	12	1.00	0.00
PSTJ	24	1.80	0.00
SOTJ	6	1.00	0.00
PSOTJ (residual)	24	1.00	1.00

than tasks 1 and 3 (five each), performance time on task 2, averaged over all other sources of variation, should be shorter. The task effect reflects this characteristic of the performance measurement (variance component for task = 20). And variation across *judges* shows whether observers are using the same criterion when timing performance. From a measurement point of view, main-effect sources of error influence absolute decisions about the

speed of performance (regardless of how other soldiers performed; called "absolute decisions"). The soldiers' performance times will depend on whether they are observed by a "fast" or "slow" timer, at a "fast" or "slow" station, and so on.

The remaining sources of variation in Table 3 reflect combinations or "statistical interactions" among the factors. Interactions between persons and other sources of error variation represent unique, unpredictable effects; the particular performance times assigned to soldiers have one or more components of unpredictability (error) in them. As a consequence, different tasks, observers, or occasions might rank order soldiers differently and unpredictably.[5] The soldier x judge effect (variance component = 8.75), for example, indicates that observers did not agree on the times they assigned to each soldier. If observer 1, for example, were used in the performance measurement, soldier 1 might be timed as faster than soldier 4. If observer 4 were used, the rank ordering would be reversed. The soldier x task interaction indicates that soldiers who performed quickly on task 1 also performed quickly on the other tasks, compared to their peers. The rank ordering of soldiers apparently does not depend on the task they performed. This is why the internal consistency coefficient, based on classical theory, was so high (.99). The soldier x occasion x judge interaction indicates judges disagreed on performance times they assigned each soldier, and the nature of this disagreement changed from one occasion to the next (negligible, Table 3). The most complex interaction, soldiers x stations x occasions x tasks x observers, reflects the effect of an extremely complex combination of error sources and other unmeasured and random error sources. It is the residual that accounts for the remaining variation in all performance times.

The remainder of the interactions do not involve persons. As a consequence, they do not affect the rank ordering of soldiers. However, they do affect the *absolute* performance-time score received by each soldier. For example, a sizable occasion x judge interaction would indicate that the performance times received by soldiers depend both on who observes them and on what occasion that observation occurs. A sizable task x judge interaction would indicate that the performance times received by soldiers depends on the particular task and observer. In doing task 1, for example, the soldiers would want judge 3 because she assigns the fastest times on this task while, in performing task 3, they might want judge 1 because he assigns the fastest times on that task.

[5] Technically, an interaction could also occur when soldiers have identical rank orders across, say, occasions *and* the distance between soldiers' performance times on each occasion is different (an ordinal interaction). An interaction with reversals in rank order (a disordinal interaction) is more dramatic and, for simplicity, is used to describe interpretations of interactions in this paper.

Improvement of Performance Measurement

Just as the Spearman-Brown prophecy formula can be used to determine the number of items needed on a test to achieve a certain level of reliability, the magnitudes of the sources of error variation can also be used to determine the number of occasions, observers, and so on that are needed to obtain some desired level of generalizability (reliability). For example, the effects involving judges (soldier x judge, judge x task, judge x task x occasion, etc.) can be used to determine whether several judges are needed and whether different judges can be used to score the performance of different soldiers, or whether the same judges must rate all soldiers due to disagreements among them. The analysis of the performance-time data in Table 3 suggests, based on the pattern of the variance component magnitudes, that several judges are needed and that the same set of judges should time all soldiers (e.g., variance components for *PJ* and *OJ*).

Generalizability of the Performance Measurement

Generalizability theory provides a summary index representing the consistency or dependability of a measurement. This coefficient, the "generalizability coefficient," is analogous to the reliability coefficient in classical theory. The coefficient for relative decisions reflects the accuracy with which soldiers have been rank ordered by the performance measurement, and is defined as:

$$\hat{\rho}^2_{Rel} = \frac{\text{Soldier Variance}}{\text{Soldier Variance} + \text{Sum of Each Interaction with Persons} / n'}$$

where n' is the number of times each source of error is sampled in an application of the measurement. For the data in Table 3, with n = 1 station, occasion, task, and judge:

$$\hat{\rho}^2_{Rel} = \frac{14.10}{14.10 + 20.30} = 0.41 .$$

The G coefficient for absolute decisions is defined as:

$$\hat{\rho}^2_{Abs} = \frac{\text{Soldier Variance}}{\text{Soldier Variance} + \text{Sum of Main Effects and All Interactions} / n'}$$

where n' is the number of times each source of error is sampled in an application of the measurement. For the data in Table 3, with n = 1 station, occasion, task, and judge:

$$\hat{\rho}^2_{Abs} = \frac{14.10}{14.10 + 54.60} = 0.20 \ .$$

Regardless of whether relative or absolute decisions are to be made on the basis of the performance measurement, the dependability of the measure based on the G theory analysis is considerably different than the analysis based on classical theory. In these examples, it is especially important to sample occasions and judges extensively for relative decisions and to sample tasks extensively as well for absolute measurements.

Summary: Revolver Test With Accuracy Scores

Recall that both time and accuracy were recorded by four observers judging soldiers' performance in the caliber .38 revolver performance test. By way of reviewing the application of G theory to performance measurements, hypothetical data on accuracy is presented. This is not merely a repeat of what has gone before. The accuracy data call for a somewhat different analysis than the performance-time data.

Design of the Revolver Test Using Accuracy Scores

In the generalizability study, each of five soldiers performed the revolver test four times: on two different occasions (*O*) at two different test stations. The soldiers' (*P*) performance on each of the three tasks (*T*) and subtasks (*S*) (see Table 1) was *independently* judged by four observers (*J*). Hypothetical accuracy scores for this G study design are presented in Table 4. The data in Table 4 have been collapsed over stations. This seemed justifiable. Because of the nature of the revolver task, stations did not introduce significant measurement error. Further, to simplify the analysis, only two of the three tasks were selected: loading and unloading/cleaning the revolver. Including the stoppage removal task would have created an "unbalanced" design, with five subtasks for tasks 1 and 3 each and only three subtasks for task 2. (See the later discussion of unbalanced designs.)

The data in Table 4 represent a soldiers x occasion x task x subtask:task x observer (*P x O x T x S:T* x *J*) design. Notice that each of the two tasks—loading and unloading—contain somewhat different subtasks. So identical subtasks do not appear with each task and we say that subtasks are nested within tasks (cf. a nested analysis of variance design). The consequence of nesting can be seen in Table 5, where not all possible combinations of *P*, *0*, *T*, *S:T*, and *J* appear in the source table as was the case in Table 3. This is because all terms that include interactions of *T* and *S:T* together cannot be estimated due to the nesting (see the later discussion of nesting).

TABLE 4 Caliber .38 Revolver Operation and Maintenance Task: Accuracy

			Observer			
Occasion	Task	Subtask	1	2	3	4
1			0	0	1	1
			0	0	1	1
		1	0	0	0	1
			1	1	1	0
			1	1	1	0
			0	0	1	1
			0	1	1	1
		2	1	1	1	1
			1	1	1	0
			1	1	1	0
			0	0	0	1
			0	0	1	1
	1	3	0	1	1	1
			0	1	1	1
			1	1	1	0
			0	0	0	1
			0	0	1	1
		4	0	0	1	1
			0	0	1	1
			0	1	1	1
			0	0	0	1
			0	0	0	1
		5	0	0	1	1
			0	1	1	1
			0	1	1	1
			* * *			
2			0	0	0	0
			0	0	0	1
		1	1	0	0	0
			0	0	1	1
			0	1	1	1
			1	0	1	0
			0	1	1	1
		2	0	1	0	1
			1	1	1	1
			1	1	1	0

			Observer			
Occasion	Task	Subtask	1	2	3	4
			1	0	1	1
			0	0	1	1
	2	3	1	1	0	1
			1	1	1	0
			1	1	1	1
			0	0	0	0
			0	0	0	1
		4	0	0	0	1
			1	1	0	0
			0	1	1	1
			0	0	0	0
			0	0	0	0
		5	0	0	0	1
			1	0	1	0
			1	1	1	1

Variance Components and G Coefficients

In G theory, interest attaches to the estimated variance components.[6] In Table 5, the variance component for soldiers, $s^2_p = .03$, reflects universe-score variance—systematic differences among soldiers that decision makers want to know about.

The remaining 22 terms in the table represent potential sources of measurement error. Relative to other components, variance components *PJ*, *PJS:T*, and *POJS:T* are sizable. Notice that in each component soldiers (*P*) and observers (*J*) are involved. Observers apparently do not agree with one another in scoring individual soldiers' performance, and this disagreement among observers changes with subtask and occasion. As a result, the *G* coefficient for a measurement made with one observer on one occasion is: .12 for relative decisions and .10 for absolute decisions.

[6] A comparison of variance components in Table 5 with variance components in Table 3 reveals substantial differences in magnitudes due to differences in metrics. Compare Tables 2 and 4.

TABLE 5 Generalizability Study for a Soldier (P) x Occasion (O) x Task (T) x Subtask:Task (S:T) x Observer (J) Design

Source of Variation	df	Mean Squares	Variance Components
Soldiers (*P*)	4	3 .140	.030
Occasions (*O*)	1	0 .010	.000
Tasks (*T*)	1	0 .810	.000
Subtasks:Tasks (*S:T*)	8	0 .881	.013
Observers (*J*)	3	2 .093	.005
PO	4	0 .123	.000
PT	4	0 .135	.000
PS:T	32	0 .153	.000
PJ	12	0 .885	.025
OT	1	0 .000	.000
OS:T	8	0 .036	.000
OJ	3	0 .837	.012
TJ	3	0 .330	.000
JS:T	24	0 .376	.014
POT	4	0 .013	.000
POS.T	32	0 .036	.000
POJ	12	0 .224	.013
PTJ	12	0 .230	.000
PJS:T	96	0 .248	.074
OTJ	3	0 .067	.000
OJS:T	24	0 .091	.000
POTJ	12	0 .071	.000
POJS:T	96	0 .100	.100

Modifications for Future Decision Studies

This pattern of findings suggests one or some combination of modifications to the performance test:

(1) Modify procedures so that observers are not also test administrators—lapses in attention may give rise to inconsistencies.
(2) Train observers more extensively and maintain training checks over the period of performance testing.
(3) Increase the number of observers judging performance.

Only the last recommended change can be evaluated with the hypothetical data. By using four observers, the G coefficients are .36 and .29 for relative and absolute decisions, respectively. Clearly, modifications in testing

procedures, training procedures, or both may be needed to increase generalizability within practical manpower and cost limits.

SKETCH OF GENERALIZABILITY THEORY

Background

Generalizability theory evolved out of the recognition that the concept of undifferentiated error in classical test theory provided too gross a characterization of the multiple sources of variation in a measurement. The multifaceted nature of error was portrayed in the last section using a hypothetical performance measurement: operating and maintaining a caliber .38 revolver. For example, a soldier might be observed by one of many observers, on one of many possible occasions, performing one of many tasks. G theory assesses each source of error—observers, occasions, tasks in this example—to characterize the dependability of the measurement and improve the job performance test design.

G theory views a behavioral measurement as a sample from a universe of admissible observations. The universe is characterized by one or more sources of error variation or "facets" (e.g., observers, occasions, tasks). This universe is typically defined as all combinations of the levels (called conditions in G theory) of the facets. In the last section's example, the universe for performance-time measurement was characterized by four facets: stations, occasions, tasks, and observers. The universe of admissible observations consisted of performance-times measured by all combinations of stations, occasions, tasks, and observers. The soldier's performance times—obtained from four observers on two tasks performed at two stations on two occasions—represented a sample from the universe of admissible observations. The decision maker intended to generalize these performance measurements to the entire universe of admissible observations.

Since different measurements may represent different universes of admissible observations, G theory speaks of the ideal datum as universe scores, rather than true scores as does classical theory, acknowledging that there are different universes to which decision makers may wish to generalize. Likewise, the theory speaks of generalizability coefficients rather than *the* reliability coefficient, realizing that the computed value of the coefficient may change as the definition of the universe changes.

Variance Components

In G theory, a measurement is decomposed into a component for the universe score (analogous to the true score in classical theory) and one or more error components (facets). To illustrate this decomposition, a two

TABLE 6 Equations for Scores and Coefficients in Generalizability Theory: S x J x O Design

	Equation	
1	Decomposition of observed score	$X_{sjo} = \mu$ (grand mean) $+\mu_s - \mu$ (soldier effect) $+\mu_j - \mu$ (judge effect) $+\mu_o - \mu$ (occasion effect) $+\mu_{sj} - \mu_s - \mu_j + \mu$ (soldier × judge effect) $+\mu_{so} - \mu_s - \mu_o + \mu$ (soldier × occasion effect) $+\mu_{jo} - \mu_j - \mu_o + \mu$ (judge × occasion effect) $+X_{sjo} - \mu_s - \mu_j - \mu_o + \mu_{sj} + \mu_{so} + \mu_{jo} - \mu$ (residual)
2	Total variance of a score	$\sigma^2 X_{sjo} = \sigma^2_s + \sigma^2_j + \sigma^2_o + \sigma^2_{sj} + \sigma^2_{so} + \sigma^2_{jo} + \sigma^2_{sjo,e}$
3	Error variance for relative decisions	$\sigma^2_{Rel} = \frac{\sigma^2_{sj}}{n'_j} + \frac{\sigma^2_{so}}{n'_o} + \frac{\sigma^2_{sjo,e}}{n'_j n'_o}$
4	Error variance for absolute decisions	$\sigma^2_{Abs} = \frac{\sigma^2_j}{n'_j} + \frac{\sigma^2_o}{n'_o} + \frac{\sigma^2_{sj}}{n'_j} + \frac{\sigma^2_{so}}{n'_o} + \frac{\sigma^2_{jo}}{n'_j n'_o} + \frac{\sigma^2_{sjo,e}}{n'_j n'_o}$
5	Generalizability coefficient for relative decisions	$E\hat{\rho}^2_{Rel} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{Rel}}$
6	Generalizability coefficient for absolute decisions	$E\hat{\rho}^2_{Abs} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{Abs}}$

facet design is used for simplicity: soldiers x judges x occasions. The object of measurement, soldiers, is not a source of error and, therefore, is not a facet. (The argument presented here readily extends to more complex designs such as those described in the previous section.)

The score X_{sjo} assigned to a particular soldier's (s) performance by a particular judge (j) on a particular occasion (o) can be decomposed into eight components as shown in equation 1 in Table 6. The single observed score X_{sjo} and each component other than mu (μ) in equation 1 has a distribution. For all soldiers in the universe of judges and occasions, the distribution of μs-μ has a mean of 0 and a variance denoted by σ^2_s (called the universe-score variance). Similarly, there are variances associated with each of the

error components: they are called variance components. For example, the variance of μ_o-μ is denoted by σ^2_o. The variance of the collection of X_{sjo} for all soldiers, judges, and occasions included in the universe, then, is the sum of all of the variance components in equation 2 (see Table 6). In words, the variance of the scores can be partitioned into independent sources of variation due to differences between soldiers, judges, occasions, and their interactions.

Numerical estimates of the variance components can be obtained from an analysis of variance (ANOVA) by setting the expected mean squares for each component equal to the observed mean squares and solving the set of simultaneous equations as shown in Table 7. (This is how the numerical values of the components of variance in Table 3 were obtained.) The variance component for soldiers (σ^2_s) represents universe-score variance, and the remaining components represent error variance. G theory focuses on the variance components. The relative magnitudes of the components provide information about particular sources of error in scores assigned to the soldiers' performance.

As an example, the performance-time data reported in Table 2 were pooled over stations and tasks to produce data in the form of a soldier x judge x occasion G study. The results of the analysis of these hypothetical data are presented in Table 8. The observed score variance is partitioned into its sources in the first column of the table. The variance components in the last column are obtained by setting the mean squares in the table equal to their expectations and solving the equations shown in Table 7. For example,

$$\sigma^2_{res} = MS_{res} = 1.20;$$

$$\sigma^2_{jo} = \left(MS_{jo} - MS_{res}\right) / n_s = (49.2 - 1.2) / 5 = 9.60;$$

and so on.

The variance component for soldiers in Table 8 represents systematic (universe-score) variation in performance times among soldiers—it is the signal the decision maker is looking for. The variance component for occasions is sizable; it represents mean differences in the times on occasions 1 and 2. Soldiers were slower on one occasion than on the other. The small variance component for judges indicates that the judges were very close in reporting average time to complete the revolver exercise. The sizable *SJ* interaction component indicates that judges disagree as to which soldiers were faster than others. The large *SO* interaction component indicates that the difference among soldiers' performance times changes from one occasion to the next; i.e., some soldiers perform faster, others slower, and this ordering changes over occasions. The relatively small *OJ* variance component indicates that judges are reasonably consistent in the mean performance

TABLE 7 Estimates of Variance Components for a Two-Facet S x J x O Design

Source of Variation	Mean Square	Expected Mean Square	Estimated Variance Component
Soldier (s)	MS_s	$\sigma^2_{io} + n_o\sigma^2_{sj} + n_j n_o \sigma^2_s$	$MS_{so} - MS_{sj} + MS_{res}) / n_j n_o$
Judge (j)	MS_j	$\sigma^2_{io} + n_o\sigma^2_{sj} + n_s n_o \sigma^2_j$	$MS_{jo} - MS_{sj} + MS_{res}) / n_s n_o$
Occasion (o)	MS_o	$\sigma^2_{io} + n_j\sigma^2_{so} + n_s n_j \sigma^2_o$	$MS_{jo} - MS_{so} + MS_{res}) / n_s n_j$
sj	MS_{sj}	σ^2_{ij}	$- MS_{res}) / n_o$
so	MS_{so}	σ^2_{io}	$- MS_{res}) / n_j$
jo	MS_{jo}	σ^2_{io}	$- MS_{res}) / n_s$
sjo,e	MS_{res}		σ^2_{es}

NOTE: n_s = number of soldiers; n_j = number of judges; n_o = number of occasions

TABLE 8 Analysis of Hypothetical Data from the Soldier x Occasion x Judge Design

Source of Variation	df	Mean Squares	Estimated Variance Components
Soldiers (*S*)	4	506.20	42.30
Occasions (*O*)	1	607.40	22.20
Judges (*J*)	3	175.40	7.35
SO	4	116.00	28.65
SJ	12	53.70	26.25
OJ	3	49.20	9.60
SOJ,*e* (residual)	12	1.20	1.20

SOURCE: Based on hypothetical data in Table 2.

times they record on the two occasions. And the small residual component indicates that other unidentified sources of error have little influence on the measurement.

Generalizability and Decision Studies

Typically enlistees are tested on only one occasion and their performance is scored by a single judge who is also responsible for administering the test. This procedure is dictated in large part by cost and convenience, and perhaps also by lack of information on the consequences this procedure has for the reliability of the performance measurement. Generalizability theory provides a method for estimating measurement error due to inconsistencies arising from one occasion to another, or from one judge to another. Once the important source of error has been identified, reliability can be forecast for alternative performance-measurement procedures. That is, with some front-end cost and inconvenience, a performance measurement program can be designed to minimize cost and inconvenience for a given level of reliability.

G theory recognizes that certain studies (G studies) are associated with the development of a measurement procedure, for example, to determine the relative influence of the sources of measurement error (judges and occasions in our example) on the dependability of a performance measurement. This information enables the test developer to recommend the number of times each potential measurement facet needs to be sampled to obtain a dependable measurement for decision making purposes. The universe of admissible observations in the G study is defined as broadly as possible within practical and theoretical constraints to estimate as many variance

components as possible. To this end, Cronbach et al. (1972) recommended using a crossed G study design so that all of the possible variance components can be estimated. By doing so, the findings will have wide applicability to as many practical measurement settings as possible.

Other studies, decision (D) studies, then apply the procedure; decisions are made, at least in part, on the basis of the measurement. For example, military policy makers might examine some aspects of readiness by measuring individual and unit performance. Or they might revise enlistment standards based on the correlation between performance and measures of these standards.

The results of the G study provide information for optimizing the design of the D study. If the G study showed that one (or more) source of error variation was very small, the decision maker could reduce the number of conditions of the facet, select (and thereby control) one level of the facet, or ignore the facet altogether. This permits a smaller—and presumably less costly—design for the decision study than that used in the G study. If, however, soldiers' performances varied by occasion, as was the case in the example G study (Table 8), soldiers should be observed on multiple occasions, and the average of their scores (performance times) should be used in further analyses (such as correlating enlistment attributes with job performance measures).

In an important sense, a D study reflects the results of applying a Spearman-Brown-like prophecy formula (in classical theory) to determine how many judges and occasions are needed for the performance measurement to meet minimal reliability (generalizability) standards. For example, the generalizability of the performance measurement in the *S* x *O* x *J* study can be increased by decreasing error associated with inconsistency across occasions—variances associated with occasions and the interactions: soldier x occasion, judge x occasion, and residual. This error variance can be reduced by taking the average of the occasion scores. This tack has the effect of reducing all variance components involving occasions by $1/n_o'$, where $n_{o'}$ is the number of occasions to be sampled in the D study. For example, suppose to reduce measurement errors due to inconsistencies among occasions, a D study were planned to take the average of three occasions' performance times on the revolver test. To determine what the variance component for occasions would be in the D study, we divide the variance component for occasions in Table 8 by three, with the result that $\sigma^2_o = 2.45$. The *SO* variance component would be reduced to 28.65/3 = 9.55, and the residual variance component would be .40. Similarly, error variance due to judges could be decreased by averaging scores over judges. The variance components reflecting variation over judges would correspondingly be divided by nj' , the number of judges to be sampled in the D study, and the residual component would be divided by the product of n_o' and $n_{j'}$.

Relative and Absolute Decisions

G theory recognizes that decision makers may use the same measurement in different ways. Some interpretations may focus on individual differences among soldiers. For example, determining the relation between enlistment attributes and military job performance depends only on the rank-ordering of soldiers on enlistment variables and on job performance variables. The decision maker, then, is concerned mainly with the generalizability of the rank ordering of soldiers over the facets of the measurement (judges and occasions in our example). We speak of *relative* decisions in this case.

Other interpretations may focus on the level of performance itself, without reference to other soldiers' performance. The written examination for a driver's license is an example—pass or fail is determined by the number of incorrect answers, not by how others performed on the examination. Likewise, decisions about military readiness might depend on absolute standards, not on how other units did. In both cases, decision makers are interested in absolute levels of performance and hence are concerned with the generalizability of absolute decisions.

Of course, decision makers can be and are interested in both kinds of decisions. The important point is that the distinction between relative and absolute decisions has important implications for the definition of measurement error and, as a consequence, the dependability of a performance measurement.

Measurement Error for Relative Decisions

For relative decisions, the error variance consists of all variance components that affect the rank-ordering of soldiers; all variance components representing interactions of the facets with the object of measurement—soldiers in our example. The error variance for relative decisions is shown in equation 3 of Table 6. The error variance for relative decisions reflects disagreements among judges and inconsistencies over occasions about the ordering of soldiers' performance. These disagreements and inconsistencies are considered error because they do not reflect systematic differences in soldiers' performance, yet they change, in unpredictable ways, measures of soldiers' performance.

Notice that the remaining score components in an observed score (see equation 1 in Table 6) are constant for all soldiers. Consequently, they do not influence the rank ordering of performance and are not defined as relative error.

Measurement Error for Absolute Decisions

For absolute decisions, the error variance consists of all variance components

except that for universe scores (see equation 4 in Table 6). The error variance for absolute decisions reflects differences in mean ratings of soldiers across judges and occasions, as well as disagreements about the ranking of soldiers' performance. When the decision maker is concerned with the absolute level of performance, the variance components associated with effects of judges and occasions (σ^2_j, σ^2_o, and σ^2_p) are defined as error variance.

The leniency of one judge as compared to another will influence a soldier's score as might, for example, the soldier's mental and physical condition on the particular day on which the test is given. That is, perceptions of particular judges who observe a soldier and the events occurring during the particular occasion will influence the observed score and, hence, the decision maker's estimate of the soldier's universe score. Thus, it may be important to obtain measures of soldiers' performance on several occasions using several judges so that these influences will be averaged out.

Generalizability Coefficients for Relative and Absolute Decisions

While stressing the importance of the variance component, G theory also provides a coefficient analogous to the reliability coefficient in classical theory. The *generalizability coefficient* for relative decisions is given in equation 5 (Table 6). In the equation, the symbol $E\rho^2$, indicates the expected value of the squared correlation between observed scores and universe scores. An analogous coefficient can be defined for absolute decisions as in equation 6 of the same table.

The generalizability coefficient, $E\rho^2$, indicates the proportion of observed-score variance ($\sigma^2_s + \sigma^2_{Rei}$) or ($\sigma^2_s + (\sigma^2_{Abs}$) that is due to universe-score variance (σ^2_s). In this respect, it is a ratio of signal to [signal + noise]. It ranges from 0 to 1.00 and, like the reliability coefficient in classical theory, its magnitude is influenced by variation among soldiers' scores and the number of observations made. The number of observations is taken into account in much the same way as in the Spearman-Brown formula in classical theory (see the discussion on D studies above). Using the Spearman-Brown formula, one can estimate the reliability of a test of any length from the reliability of the original test. Analogously, in equations 3 and 4 (Table 6), the denominator indicates the number of observations to be made in the D study (number of judges and occasions). As the number of observations increases, the error variance (σ^2_{Rei}) or (σ^2_{Abs}) decreases and the generalizability coefficient, $E\rho^2$, increases.

A major contribution of generalizability theory is that it allows the researcher to pinpoint the sources of measurement error (e.g., judge, occasion, or both) and to increase the appropriate number of observations so that error "averages out." The researcher can estimate how many conditions of each facet are needed to obtain a certain level of generalizability. If, for example, variation due to occasions is large relative to variation due to soldiers, and

variation due to judges is small, increasing the number of occasions would produce a lower estimate of error variation and consequently a higher generalizability coefficient (see equation 4, Table 6), whereas increasing the number of judges would have little effect on the estimates of error variation and generalizability. (See Shavelson and Webb, 1981, for details on optimizing generalizability coefficients by selecting conditions of two facets.)

Random and Fixed Facets

To this point, the presentation of G theory has assumed that the conditions of the measurement have been randomly sampled from an indefinitely large universe.[7] In the hypothetical S x J x O design, we assumed that the four judges were sampled randomly from a universe of judges, and that the two occasions were sampled randomly from a universe of admissible occasions. In the previous section, a more comprehensive design was described: soldiers x stations x occasions x tasks x observers. The analysis assumed that the two stations were a random sample from an indefinitely large universe, that the two occasions were a random sample, that the three tasks were a random sample from a large universe of tasks, and that the observers were sampled randomly.

Generalizability theory, then, can model the military decision maker's ideal performance measurement. This is a measurement that generalizes over all possible stations at which the test might be given, over all possible occasions on which the test might be given, over all possible tasks in a military occupational specialty, and over all possible observers who might time and score soldiers' performance.

The assumption of random facets is more an ideal than an actuality in performance measurement. The truth be known, in most performance measurement studies stations are not randomly sampled, occasions are not randomly sampled, tasks are not randomly sampled, and judges are not randomly sampled. Indeed, soldiers may not be randomly sampled. This fact is made clear in reports from all four Services [Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics), 1983]:

- It is simply not feasible to develop a measure of performance for every task performed by a soldier in an MOS. . . . The variables considered important in selecting tasks for testing were frequency, criticality, and variability of performance (p. 25).
- These tests will be developed for tasks or part tasks that are (1) routinely done by at least 90% of first-term sailors . ., (2) feasible to test in a

[7] Variance component estimates for restricted universes are readily available (see Brennan, 1983; Cardinet and Allal, 1983).

hands-on mode on actual equipment, and (3) not severely affected by limited operational constraints (i.e., equipment only partially operable) (p. 58).

- From occupational survey data, [AF] critical specialty and job-type specific tasks will be identified. Subject matter experts (SMEs) will aid the developers in dichotomizing the tasks into those which can be economically observed and those which must be measured by interview. The SME's will then develop procedures for observing task performance . . . (p. 45)
- [Marine Corps] test administrators will be trained in how to give and score hands-on tests . . . . The administrators will be assigned full-time to the research project (p. 73).

Generalizability theory handles sampling issues in two ways. The first way is to assume that the conditions of a measurement facet are exchangeable with other potential conditions (see Shavelson and Webb, 1981, for details). Thus, if the test administrators (observers or judges) are considered exchangeable for other observers who might have been used, or if the test occasions are considered exchangeable for other occasions, these facets might legitimately be considered random. In this case, G theory treats these facets as random and proceeds with an analysis like that presented in Table 8.

A second way G theory treats these sampling issues is to recognize that either (1) decision makers are not interested in generalizing beyond the conditions of the facet, or (2) the conditions of the facet exhaust the universe of possible conditions (cf. a fixed factor in the ANOVA). Hence, the sample of tasks, n(tasks), equals the universe of tasks, N(tasks). For example, the three tasks comprising the caliber .38 revolver operation and maintenance test—loading, reducing a stoppage, and unloading/cleaning the weapon—might exhaust the universe of tasks over which decision makers might wish to generalize. In this case, the task facet would be considered a "fixed facet."

Statistically, G theory treats fixed facets by averaging over the conditions of a fixed facet and examining the generalizability of these averages over the random facets (Cronbach et al., 1972:60; see Erlich and Shavelson, 1976, for a proof). This tack is justified because by averaging over the conditions of a fixed facet the average is over the entire universe. Averaging provides the best score for an individual because it represents the individual's universe score over the conditions of the fixed facet. The statistical analysis of data from G studies with a fixed facet proceeds by carrying out an ANOVA on scores averaged over the conditions of the fixed facet.

Consider a modification of the hypothetical *S* x *J* x *O* generalizability study (Table 8). Suppose that the judges observed the speed of the soldiers' performances on three tasks—loading, reducing a stoppage, and unloading—and that these three tasks exhausted the universe of possible revolver tasks. This modification produces a three-facet G study with two random

facets (occasions and judges) and one fixed facet (tasks). Each soldier's performance-time data would be averaged over the three tasks for each judge and each occasion. An *S* x *J* x *O* ANOVA would be used to analyze these averages.

An analysis of hypothetical performance-time data from the *S* x *J* x *O* x *T** generalizability study (where *T** denotes a fixed facet) is provided in Table 9. Two things about the table are noteworthy. First, the results in part (a) are similar to those reported in the previous section for the soldier x station x judge x occasion x task—completely random design, since the variance components are (virtually) uncorrelated and, anyway, the effects of the station facet and its interactions with all other facets were 0.

Second, the data were analyzed as a completely random design and mean squares and variance components were recomputed for the restricted universe of generalization—to occasions and judges, weighted for the number of tasks (Table 9, part b). All of the components involving task variation (e.g., *T*, *ST*, *OT*, *JT*, *SOT*)—the within-T components—go to 0 because scores are averaged over facet *T* (see Cronbach et al., 1972:115). This is the same result that would have been obtained if a *S* x *J* x *O* design had been run on performance times averaged over tasks.

Incidentally, the generalizability coefficients for relative decisions are the same for the two models (Table 9, part c). This happened because, in the *S* x *O* x *J* x *T* (random) model, the person x task and higher-order person x task x (etc.) variance components are 0 except in the case where $\sigma^2_{sot} = 1.00$. Compared to the other sources of relative error (e.g., *SO* and *SJ*), the person x task interactions have no appreciable effect on the relative G coefficient (with "real" data, there is no a priori reason to expect soldier x task variance components to be 0).

This is not the case for the absolute G coefficient where the variance component for task enters into the definition of measurement error in the random model and is sizable (20). Even after dividing it by the number of tasks sampled (3), it is still sizable (6.67).

Finally, the results of the *S* x *O* x *J* x *T** G study reported in Table 9 are identical to those in Table 8 that reported the data from a *S* x *J* x *O* design. In both analyses, the task facet is ignored (averaged over) as is the station facet.

Before leaving the topic of fixed facets, we should note that the definition of a fixed facet that leads one to "average over" its conditions might be misleading in certain circumstances (see Shavelson and Webb, 1981). In a teaching context, for example, it might not make sense to average over elementary teachers' performance in teaching mathematics *and* English. Teachers do different things in teaching the two subject matters, and to ignore these differences by averaging over them would lead to misconceptions and loss of information. A redefinition of the fixed facet that recognizes such differences,

TABLE 9 Analysis of Data from a S x O x T x J* Design Where T* is a Fixed Facet

Source of Variation	df	Mean Square	Variance Components[a]
(a) *Random Model*			
Soldier (*S*)	4	1020.80	14.10
Occasion (*O*)	1	1273.00	7.40
Task (*T*)	2	1659.00	20.00
Judge (*J*)	3	349.80	2.45
SO	4	239.00	9.55
ST	8	9.80	0.00
SJ	12	106.80	8.75
OT	2	59.80	1.25
OJ	3	97.80	3.20
TJ	6	1.80	0.00
SOT	8	9.80	1.00
SOJ	12	1.80	0.00
STJ	24	1.80	0.00
OTJ	6	1.80	0.00
SOTJ,e (residual)	24	0.40	0.40
(b) *Mixed Model with Task Fixed*			
S	4	506.20	42.30
O	1	607.40	22.20
J	3	175.40	7.35
SO	4	116.00	28.65
SJ	12	53.70	26.25
OJ	3	49.20	9.60
SOJ,e (residual)	12	1.20	1.20
(c) *Generalizability Coefficients*[b]			
Random Model		Mixed Model (*T** Fixed)	
$(REL) = \frac{14.10}{14.10 + 7.15} = 0.66$		$\frac{42.30}{42.30 + 21.04} = 0.67$	
$(ABS) = \frac{14.10}{14.10 + 18.73} = 0.43$		$\frac{42.30}{42.30 + 35.18} = 0.55$	

NOTES: [a] The expected mean squares equations for the mixed design with T* fixed are:

$$E[(MS)S|T^*] = n_t\sigma^2_{Res} + n_t n_j\sigma^2_{so} + n_t n_o\sigma^2_{sj} + n_t n_o n_j\sigma^2_s$$

$$E[(MS)O|T^*] = n_t\sigma^2_{Res} + n_t n_s\sigma^2_{oj} + n_t n_j\sigma^2_{so} + n_t n_s n_j\sigma^2_o$$

$$E[(MS)J|T^*] = n_t\sigma^2_{Res} + n_t n_s\sigma^2_{oj} + n_t n_o\sigma^2_{sj} + n_t n_s n_o\sigma^2_j$$

$$E[(MS)SO|T^*] = n_t\sigma^2_{Res} + n_t n_j\sigma^2_{so}$$

$$E[(MS)SJ|T^*] = n_t\sigma^2_{Res} + n_t n_o\sigma^2_{sj}$$

$$E[(MS)OJ|T^*] = n_t\sigma^2_{Res} + n_t n_s\sigma^2_{oj}$$

$$E[(MS)RES|T^*] = n_t\sigma^2_{Res}$$

[b] The G coefficients are calculated for the G study with two occasions, three tasks (random model), and four observers.

then, might be appropriate. There may be analogous situations in military performance testing, such as distinguishing between daytime and nighttime performance. If there is any doubt, the analysis of the fixed facet would proceed in two stages. In the first stage, a G study analysis would be carried out treating all facets as random to assess the variability among conditions of the (candidate) fixed facet. Large variance components associated with the candidate facet would suggest that performance differs across conditions; a large person by candidate-facet interaction would indicate that individuals' performances are not ordered the same under different conditions of the candidate facet. If this is the case, a second stage might be to conduct G study analyses for each condition of the candidate facet separately.

G and D Studies With Crossed and Nested Facets

In the previous examples, *crossed* designs have been used to illustrate generalizability theory. In a crossed design, the levels of each variable are combined with the levels of all other variables. In the example G study (Table 8), all soldiers were observed by all judges on both testing occasions. We denoted this crossed design by the notation, $S \times J \times O$, indicating there are $n_s \times n_o \times n_j$ (= 5 x 2 x 4) = 40 combinations of conditions in the design.

There are cases where not all conditions of one facet can be combined

with all conditions of one or more other facets. One example is the G study design in which soldiers' performance was scored on two occasions by four judges on the subtasks of loading and unloading/checking a caliber .38 revolver (see Table 4). The subtasks involved in loading the revolver are not exactly the same as those involved in unloading the revolver. That is, the subtasks are unique to the particular task performed. We say that subtasks are nested within tasks and write the design symbolically to represent this nesting: *P* x *O* x *J* x *T* x *S:T* where *S:T* is read, "subtasks are nested within tasks," and *P* refers to soldiers (mnemonically, persons).

Nesting also arises in traditional achievement testing, where the first 10 items might deal with reading comprehension and the next 10 items might examine science knowledge. We say that items ("subtasks") are nested within subject-matter topics (tasks). A design in which 100 students (S) take an achievement test with multiple topics (T) and each topic contains a different set of items (*i*) would be designated as follows: *S* x *T* x *I:T*.

A third example, common with performance measurements, is when judges A and B score the performance of soldier 1, observers C and D score the performance of soldier 2, and so on. In this case, we say that judges (*J*) are nested within soldiers (S). This design would be represented as: *S* x *J:S*. More generally, the conditions of facet A are said to be nested in the conditions of facet B when, for each condition of facet B, there is more than one condition of facet A, *and* the conditions of facet A are different for each condition of facet B.

The major consequence of nesting is that not all variance components can be estimated. In the *P* x *O* x *J* x *T* x *S:T* design, variance components for the following interactions cannot be estimated: *TS:T*, *PTS:T*, *OTS:T*, *JTS:T*, *POTS:T*, *PJT:ST*, and *OJTS:T* (see Table 4). In the *S* x *T* x *I:T* design, variance components for *TI:T* and *STI:T* cannot be estimated. And finally, the variance component for *SJ:S* cannot be estimated in the *S* x *J:S* design. In general, interaction components of variance cannot be estimated for interactions containing the nested variable and the variable in which it is nested.

Since nested designs do not provide variance components for all sources of measurement error, Cronbach et al. (1972) recommended that G studies employ crossed designs. Only when many conditions need to be sampled are nested G studies recommended.

The crossed-design recommendation, however, does not extend to D studies. At the time a D study is conducted, sources of measurement error should have already been pinpointed. The aim of the D study is to obtain large samples of conditions for errorsome facets. To this end, nesting is a boon. The hypothetical *S* x *J* x *O* generalizability study with two occasions and four judges had revealed large variance components for occasions and the interaction of soldiers and occasions (Table 8). A D study should use as many occasions as possible, but three per soldier seems like the limit. So,

the following nested D-study design might be used: *S* x *J:P* x *O* where soldiers are crossed with judges and occasions are nested within each soldier. In this example, the variance component for occasions would be sampled n_p x n_o = 5 x 3 = 15 times, not just 2 times as in the G study. And the variance component for the *JO:P* interaction, for example, would be sampled n_j x (n_p x n_o) = 60 times!

Multivariate Generalizability Theory

Performance measurements often describe performance with more than one score. In an example used in this section, two judges timed and scored the performance of five soldiers on three tasks—loading, reducing a stoppage, and clearing a revolver—on each of two occasions. Two different sets of scores may be of concern to the decision maker. The first is a profile or composite consisting of performance time and accuracy. The second is a profile or composite of performance times across the three tasks, or a profile or composite of accuracy scores across the tasks.

In assessing the reliability or generalizability of profiles or composites, most studies take a univariate approach. Tasks, for example, would be treated as a facet—source of error—in the measurement and a univariate G analysis would be conducted as has been done in the examples up to this point. Or the time and accuracy measures, for example, would be treated separately in univariate G studies, producing two different sets of findings and, perhaps, conflicting D-study recommendations. In short, the univariate approach does not assess sources of error covariation (correlation) among the multiple scores. Such information is important for designing an optimal decision study and for permitting a decision maker to determine the composite (across tasks or across time and accuracy) with maximum generalizability.

Generalizability theory provides a method for taking into account the covariances among performance-measurement scores. Just as univariate generalizability theory stresses interpretations of the pattern of variance components, multivariate G theory stresses interpretation of variance and covariance components (see Webb et al., 1983, for a concise, elementary presentation of the theory). It also provides a summary index for a composite of scores, a multivariate generalizability coefficient analogous to the univariate coefficient.

To make the presentation of multivariate G theory concrete, let's simplify the design of the revolver operation and maintenance measure to a soldier x occasion design. Three measures are obtained: scores on *time* to (1) load, (2) remove a stoppage from, and (3) unload the revolver. This design may be described as a *S* x *O* design with a set of three scores (or dependent variables).

Perhaps the easiest way to explain the multivariate version is by analogy

to the univariate case. In the univariate case, we treat the data from the *S* x *0* design with three task scores as a *S* x *0* x *T* design with one performance score per cell of the design. An observed score (e.g., time) is decomposed into the universe score and error scores corresponding to occasions, tasks, their interactions with each other and with soldiers. An estimate of each component of variation in the observed scores is obtained. For this two-facet, univariate design, σ^2_s is the estimated universe-score variance. For relative decisions, the estimate of the multifaceted error variance is:

$$\hat{\sigma}^2_{Rel} = \frac{\hat{\sigma}^2_{so}}{n'_o} + \frac{\hat{\sigma}^2_{st}}{n'_t} + \frac{\hat{\sigma}^2_{res}}{n'_o n'_t}$$

where n_o' and nt' are the numbers of conditions of the facets in the decision study, and the generalizability coefficient is:

$$\hat{\rho}^2 = \frac{\sigma^2_s}{\hat{\sigma}^2_s + \hat{\sigma}^2_{Rel}}$$

In extending the notion of multifaceted error variance to multivariate designs, we treat tasks not as a facet but as three dependent variables: time to load ($_lX_{so}$), time to remove a stoppage ($_rX_{so}$), and time to unload ($_uX_{so}$). For each measure, the components of the observed score variance reflect variation between soldiers σ^2_s and a residual, which is the interaction of students with occasions confounded with error ($\sigma^2_{so,\,e}$):

$$\sigma^2({}_lX_{so}) = \sigma^2({}_ls) + \sigma^2({}_lso,e)$$
$$\sigma^2({}_rX_{so}) = \sigma^2({}_rs) + \sigma^2({}_lso,e)$$
$$\sigma^2({}_uX_{so}) = \sigma^2({}_us) + \sigma^2({}_uso,e)$$

where

l is the loading task
r is the removing task
u is the unloading task
X is an observed score
s is soldiers
o is occasions.

So, $\sigma^2(_lX_{so})$ is the observed score variance on the loading task and $\sigma^2(^uso_{,\,e})$ is the residual variance on the unloading task. Moreover, the components of

one score (e.g., $({}_{l}X_{so})$ can be related to (covary or correlate with) the components of the other scores (e.g., $({}_{r}X_{so})$. For a *composite* of the three scores, the expected observed-score variance (the universe-score variance plus the residual variance) depends on the components of *covariance* as well as on the components of variance. The observed-score variance of this composite can be expressed as the variance-covariance matrix shown in part (a) of Table 10.

In univariate G theory, the expected observed-score variance can be decomposed into components for universe-score variance and error variances. In multivariate G theory, the expected observed-score variance-covariance matrix can also be decomposed. For relative decisions, the decomposition is given in part (b) of Table 10.

Just as the analysis of variance can be used to obtain estimated components of variance, Multivariate ANalysis Of VAriance (MANOVA) provides a computational procedure for obtaining estimated components of variance and covariance. While ANOVA provides scalar values for the sums of squares and mean squares, MANOVA provides matrices of sums of squares and crossproducts and mean squares and crossproducts.

Estimates of components of covariance are obtained by setting the expected mean product (MP) equations equal to the observed mean products and solving the set of simultaneous equations. (As in the univariate case, estimated variance components are obtained by setting the expected mean square equations equal to the observed means squares and solving the set of simultaneous equations.) The equations in part (c) of Table 10 relate mean products to their expectations such that components of variance and covariance for the universe score matrix (Table 10, part (b)) can be obtained. The first three equations in part (c) reflect the univariate case, in which each subtest is examined separately, whereas all six equations represent the multivariate case.

The results of a multivariate analysis—the variance and covariance components for persons, occasions, and residual—are given in Table 11 using hypothetical data. Part (1) contains components of universe-score variance and covariance. Universe-score variance components are found along the main diagonal, and covariance components are given off the diagonal. The high covariance components among the universe scores across the three tasks, relative to the residual covariance components, indicate that soldiers who load the revolver quickly also remove the stoppage and unload the revolver quickly. That is, there is an underlying speed component in tasks that involve operating and maintaining a caliber .38 revolver. The one high residual component of covariance suggests that unexplained factors undermine the consistency of performance times and that the loading and unloading task scores tend to fluctuate together.

The multivariate G coefficient is an analog to the univariate G coefficient

TABLE 10 Decomposition of a Composite of Scores on Three Performance Tasks-Loading, Removing a Stoppage from, and Unloading a Revolver

(a) *Observed-Score Covariance Matrix*

$$\begin{array}{lll} \sigma^2({}_lX_{SO}) & \sigma({}_lX_{SO}, {}_rX_{SO}) & \sigma({}_lX_{SO}, {}_uX_{SO}) \\ \sigma({}_lX_{SO}, {}_rX_{SO}) & \sigma^2({}_rX_{SO}) & \sigma({}_rX_{SO}, {}_uX_{SO}) \\ \sigma({}_lX_{SO}, {}_uX_{SO}) & \sigma({}_rX_{SO}, {}_uX_{SO}) & \sigma^2({}_uX_{SO}) \end{array}$$

where the variances for the loading, removing, and unloading tasks are found on the main diagonal (respectively) and $({}_lX_{SO}, {}_rX_{SO})$ is the covariance of loading-task scores with removal-task scores. $({}_lX_{SO}, {}_uX_{SO})$ is the covariance of loading-with unloading-task scores, and $({}_rX_{SO}, {}_uX_{SO})$ is the covariance of removing-and unloading-task scores.

(b) *Decomposition of the Observed-Score Covariance Matrix*

Universe-Score Variance and Covariance Components:

$$\begin{array}{lll} \sigma^2({}_lS) & \sigma({}_lS, {}_rS) & \sigma({}_lS, {}_uS) \\ \sigma({}_lS, {}_rS) & \sigma^2({}_rS) & \sigma({}_rS, {}_uS) \\ \sigma({}_lS, {}_uS) & \sigma({}_rS, {}_uS) & \sigma^2({}_uS) \end{array}$$

Residual Variance and Covariance Components:

$$\begin{array}{lll} \sigma^2({}_lRes) & \sigma({}_lRes, {}_rRes) & \sigma({}_lRes, {}_uRes) \\ \sigma({}_lRes, {}_rRes) & \sigma^2({}_rRes) & \sigma({}_rRes, {}_uRes) \\ \sigma({}_lRes, {}_uRes) & \sigma({}_rRes, {}_uRes) & \sigma^2({}_uRes) \end{array}$$

(c) *Expected Mean Squares and Products*

$$E[MS({}_lX_{SO})] = \sigma^2({}_lRes) + n_O\sigma^2({}_lS)$$
$$E[MS({}_rX_{SO})] = \sigma^2({}_rRes) + n_O\sigma^2({}_rS)$$
$$E[MS({}_uX_{SO})] = \sigma^2({}_uRes) + n_O\sigma^2({}_uS)$$

$$E[MP({}_lX_{SO}, {}_rX_{SO})] = \sigma({}_lRes, {}_rRes) + n_O\sigma({}_lS, {}_rS)$$
$$E[MP({}_lX_{SO}, {}_uX_{SO})] = \sigma({}_lRes, {}_rRes) + n_O\sigma({}_lS, {}_uS)$$
$$E[MP({}_rX_{SO}, {}_uX_{SO})] = \sigma({}_lRes, {}_rRes) + n_O\sigma({}_rS, {}_uS)$$

(d) *Multivariate G Coefficient for Relative Decisions*

$$\rho^2 = \frac{\underset{\sim}{a}' V_s \underset{\sim}{a}}{\underset{\sim}{a}' V_s \underset{\sim}{a} + \dfrac{\underset{\sim}{a}' V_s \underset{\sim}{a}'}{n'_o}}$$

where V is a matrix of variance and covariance components estimated from the mean products matrices, n' is the number of conditions of facet O in a D study, and a is the vector of canonical coefficients that maximizes the ratio of universe-score variation to universe-score plus error variation.

TABLE 11 Estimated Variance and Covariance Components Underlying the Expected Observed-Score Variance in the Soldier x Occasion Multivariate G Study of Performance on Three Revolver Tasks no =1

	Task		
Source of Variation	(1) Loading	(2) Removing	(3) Unloading
Soldiers (S)			
Loading	4.27		
Removing	1.07	4.05	
Unloading	2.08	1.14	5.84
Residual (*SO*,*e*)			
Loading	2.34		
Removing	0.00	0.78	
Unloading	0.84	0.18	1.10

(Table 10, part (d); see Shavelson and Webb, 1981; Webb et al., 1983). From a random effects MANOVA, the canonical variates (weights applied to the three task times) are determined to maximize the ratio of universe score variation to universe score plus error variation. By definition, the first composite that emerges from the analysis is the most reliable. In our hypothetical example, the multivariate generalizability of a composite composed of the three tasks with performance times obtained at one occasion is .73 for the first canonical variate.

In forming the composite to optimize the multivariate G coefficient, time to remove a stoppage was given the greatest weight (.37), while the other two tasks received considerably less weight (time to load, .11; and time to unload, .07). This weighting might not fit with what experts consider to be the correct weighting, because the composite is formed solely on statistical grounds, not on conceptual or practical ground.

A composite can be formed on conceptual or practical grounds simply by having the decision maker supply weights for the three tasks, and then by (1) taking the weighted sum of the three scores for each soldier, and (2) running an *S* x *O* univariate G analysis on the composite. The resulting univariate G coefficient for this composite can be compared with the statistically optimum multivariate coefficient to determine the precision lost to achieve conceptual or practical validity.

Symmetry in Behavioral Measurements

Typically, most behavioral measurement have been used to differentiate individuals. Hence, national achievement tests are designed to spread out individuals' scores by including difficult items. Military performance tests have been or will be designed to selectively sample tasks that differentiate individuals' military occupational specialty performance. The work of Cardinet. Tourneur, and Allal (Cardinet et al., 1976a, 1976b, 1981; Cardinet and Allal, 1983; Cardinet and Tourneur, 1974, 1977; Tourneur, 1978; Tourneur and Cardinet, 1979; for a different perspective, see Wittman, 1985) recognized, however, that the focus of measurement may change, depending on a particular decision maker's purpose. For example, the decision maker might want to know whether there are systematic differences in the behavior of different subgroups of people in a population or whether the population can be considered homogeneous. Or, the decision maker might be interested in performance scores on a set of tasks (e.g., task difficulty) and not in the differentiation of individuals' performance on the set of tasks. Hence, the task, and not the individual, becomes the object of measurement. Variation among individuals' performance on the task becomes a source of measurement error.

The principle of symmetry of behavioral measurements states that each of the facets of a factorial design can be selected as the object of study; individuals are not necessarily the sole object of measurement. This principle led Cardinet and his colleagues to distinguish among four stages of a measurement study:

(1) Observation design—choice of facets and their conditions, and the computation of mean squares;
(2) Estimation—decisions about whether the facets are finite or infinite, random or fixed, and the estimation of variance components;
(3) Measurement—specification of the facet (or combination of facets) that is the focus of measurement, and specification of the sources of error; and
(4) D-study designs.

The notion that symmetry leads to the possibility of multifaceted populations is something not considered in classical theory. This is particularly germane to military performance testing because individuals are parts of small units (e.g., crews) which are parts of larger units (e.g., platoons), which are parts of still larger units (e.g., companies), and so on. The level and consistency of an individual's performance may be a function of his ability, but it may also be a function of the training and leadership provided by the organization in which he resides. In this case, universe-score variance

might include, in addition to systematic variation among individuals, systematic variation among higher units in which the individual is embedded (see the next section; also Kahan et al., 1985).

Perhaps most importantly for military applications, the principle of symmetry provides one possible means for estimating the generalizability of unit performance, a topic beyond the scope of the Services' concerns at this point, but something that certainly should be on a future measurement agenda. For this reason, this paper on G theory has been subtitled, "I. Individual Performance." An example of the application of G theory to unit performance measurements is given in the next section.

ILLUSTRATIVE APPLICATIONS OF GENERALIZABILITY THEORY

To this point, generalizability theory has been applied to military performance measurements using hypothetical data. In this section, two published G studies are presented. These studies were selected because they illustrate the concepts and procedures described in previous sections. One is a study of the generalizability of general educational development ratings based on job descriptions found in the *Dictionary of Occupational Titles* (Shavelson and Webb, 1981; Webb and Shavelson, 1981; Webb et al., 1981). The other is a study of the generalizability of a tank-crew performance measurements (Kahan et al., 1985).

Generalizability of General Educational Development Ratings

The U.S. Department of Labor (1972) developed the General Educational Development scale to rate the amount of reasoning, mathematics, and language abilities needed to perform various jobs. General educational developmental ratings are used in several employment and training situations. For example, they provide the basis for: (1) estimates of time required to learn job skills; (2) state employment agencies' decisions to refer persons to specific employers, job training programs, or remedial education programs; and (3) equating jobs that have similar educational requirements.

G-Study Design

In this study, job analysts were given written descriptions of jobs, published in the *Dictionary of Occupational Titles*, and were asked to rate the jobs on three components of the general educational development scale: reasoning development, mathematics development, and language development. Each component was measured on a 6-point scale. Each of 71 raters

from 11 geographic field centers across the United States evaluated the three components of a sample of jobs on two occasions. Different centers had different numbers of job analysts, ranging from 2 to 12. Hence, the G study design was a partially nested, unbalanced design with different numbers of raters nested within centers. To illustrate G theory in its basic form, a random sample of two raters from each center was taken to form a balanced design. (The unbalanced design is discussed below.) Because concern attaches to estimating the general educational development required to perform each job, *jobs* is the object of measurement and the variance component for jobs (σ^2_j) is interpreted as the universe-score variance. All other variance components were considered measurement error in this study since absolute decisions are made regarding general educational development requirements for each job. These include the components for raters nested within center, center, occasion, and all interactions.

Univariate G Analysis

For the univariate G study, a random effects analysis of variance was used to estimate the variance components contributing to the observed variation in job ratings. A separate analysis was performed for each component of the general educational development scale. The results are presented in Table 12.

Since the analysis focuses on absolute decisions, the error variance, $\sigma^2_{Abs,}$ reflects not only disagreements about the ordering of jobs but also mean differences in ratings. It is important to know, for example, whether raters use essentially the same mean level of the rating scale as well as whether they rank order jobs similarly.

The estimated variance components for jobs differ across general educational development ratings. They suggest that jobs can be distinguished more on their demands for language than on their demands for mathematics and reasoning. The patterns of variance components' contribution to error were consistent: raters' ratings accounted for most of the error variation and occasions and centers accounted for little. This pattern suggested that, by taking the average rating of four raters, measurement error can be reduced by about 75 percent ($\hat{\sigma}^2_{Abs}$) in Table 12) and the G coefficients ($\hat{\rho}^2$) correspondingly increased to .87 for reasoning, .79 for mathematics, and .85 for language.

The consistent pattern of results for the ability components was not unexpected since their correlations were as follows: for reasoning and mathematics, $r = .73$; for reasoning and language, $r = .84$; and for mathematics and language, $r = .73$. The magnitude of these correlations suggests that all three general educational development ratings share a common, underlying factor and that a multivariate G coefficient would be appropriate.

TABLE 12 Univariate Generalizability Study of General Educational Development Ratings

	Estimated Variance Component			Estimated Variance Component with $n_r = 4$, $n_o = 1$, and $n_c = 1$		
Source of Variation	Reasoning	Math	Language	Reasoning	Math	Language
Jobs(*J*)	0.74	0.63	1.01	0.74	0.63	1.01
Occasions (*O*)	0.00	0.00	0.00	0.00	0.00	0.00
Centers(*C*)	0.00	0.02	0.05	0.00	0.02	0.05
Raters:C (*R:C*)	0.06	0.02	0.00	0.02	0.00	0.00
JO	0.00	0.01	0.01	0.00	0.01	0.01
JC	0.00	0.00	0.00	0.00	0.00	0.00
JR:C	0.13	0.16	0.14	0.03	0.04	0.04
OC	0.01	0.00	0.00	0.00	0.00	0.00
OR:C	0.00	0.09	0.07	0.00	0.02	0.02
JOC	0.00	0.02	0.01	0.00	0.02	0.01
JOR:C	0.22	0.25	0.22	0.06	0.06	0.05
ρ^2_{Rei}	0.42	0.57	0.50	0.11	0.17	0.18
ρ^2_{Abs}	0.64	0.53	0.67	0.87	0.79	0.85

NOTE: The design is raters nested within centers crossed with jobs and occasions. From Webb, Shavelson, Shea, and Morello (1981:190).

Multivariate G Analysis

For the multivariate generalizability study, a random effects MANOVA was run using the reasoning, mathematics, and language ratings as a vector of scores. Due to the limited capacity of computer programs available to perform the multivariate analysis and because geographic center contributed little to variability among job ratings, geographic center was excluded from the multivariate analysis. The design, then, was raters crossed with jobs and occasions with three ability scores.

For each source of variation in the design, variance and covariance component matrices were computed from the mean product matrices. Hence, one matrix, for example, contained estimated universe-score variances and covariances. All matrices with negative estimated variance components (diagonal values) were set equal to 0 in further estimation. For this analysis, the matrices of variance components, coefficients of generalizability, and canonical weights corresponding to each coefficient of generalizability were computed.

The estimated variance and covariance component matrices are presented in Table 13. Only the components for one rater and one occasion are

TABLE 13 Estimated Variance and Covariance Components for Multivariate Generalizability Study of General Educational Development Ratings (nr=1, no=1)

Source of Variation	Reasoning	Mathematics	Language
Jobs (*J*)	0.75		
	0.64	0.66	
	0.88	0.74	1.09
Occasions (*O*)	0.00		
	0.00	0.00	
	0.00	0.00	0.00
Raters (*R*)	0.03		
	0.03	0.09	
	0.03	0.05	0.05
JO	0.00		
	0.00	0.00	
	0.00	0.00	0.00
JR	0.12		
	0.11	0.13	
	0.09	0.07	0.11
OR	0.00		
	0.01	0.01	
	0.00	0.00	0.01
JRO,*e*	0.21		
(residual)	0.07	0.29	
	0.11	0.10	0.26

NOTE: The design is raters crossed with jobs and occasions with scores on reasoning, mathematics, and language. From Shavelson and Webb (1981:18).

included. To obtain results for four raters, the components corresponding to the rater effect and rater interactions need only to be divided by four.

As a consequence of the calculation procedure, the variance components in Table 13 are the same as those produced by the univariate analysis. The components of covariance, however, provide new information. The large components for jobs reflects the underlying correlations among the general educational development components. Jobs that require high reasoning ability are seen by raters to require high mathematics and language ability. Whereas the nonzero components of variance for raters indicate that some raters give higher ratings than others, the positive components of covariance

indicate that the raters who give higher ratings on one general educational development component are likely to give higher ratings on the other general educational development components. The positive components for the job x rater interaction suggest that not only do raters disagree about which jobs require more ability but their disagreement is consistent across general educational development components. The nonzero components for error suggest that the unexplained factors that contribute to the variation of ratings also contribute to the covariation between ratings. As expected, the components of covariance due to the occasion main effect and interactions are negligible.

Composites of general educational development that have maximum generalizability are presented in Table 14. When multivariate generalizability was estimated for one rater and one occasion, one dimension with a coefficient exceeding .50 emerged. This dimension is a verbal composite of reasoning and language. The analysis with four raters and one occasion produced two dimensions with coefficients over .50. One composite was defined by reasoning and language (coefficients .74 and .92 for one and four raters, respectively); the other by a mathematics-language contrast with a G coefficient of .62 for four raters and one occasion.

Unbalanced Designs

The original G study design was unbalanced with a different number of raters nested within geographic centers. The results of the unbalanced design were compared with those of two balanced designs: (1) raters randomly sampled from each of 5 randomly sampled centers, and (2) two raters randomly selected from each of the 11 centers. The estimates of variance components

TABLE 14 Canonical Variates for Multivariate Generalizability Study of General Educational Development Ratings

	Canonical Coefficients		
General Educational	n_r=1, n_o=1	n_r=4, n_o=1	
Development Component	I	I	II
Reasoning	0.34	0.38	0.05
Mathematics	0.06	0.06	-1.95
Language	0.51	0.57	1.33
Coefficient of generalizability	0.74	0.92	0.62

NOTE: The design is raters crossed with jobs and occasions.

in the unbalanced design were obtained with a modification of Rao's MIVQUE procedure suggested by Hartley et al. (1978; see Shavelson and Webb, 1981).

The magnitudes and pattern of estimated variance components from the three analyses were very similar. The generalizability coefficients ranged from .65 (unbalanced design) to .63 (two raters, five centers). These findings are consistent with those of Bell (1985). Random deletion of conditions to create a balanced design appears not to distort G study findings.

Generalizability of Unit-Performance Measurements

This G study of tank crews illustrates the application of multilevel G theory (Cronbach, 1976) and the principle of symmetry (see the earlier discussion) to hierarchical populations, a characteristic of military performance measurements. Tank-crew performance data were collected in the spring of 1971 during the annual qualification firing exercises at the Seventh Army Training Center, Grafenwohr, Germany. The tank crews performed the Table VIII mission—deliberate attack (live fire)—based on the Army Training and Evaluation Program for Mechanized Infantry Tank Force (ARTEP 71-2, 1978). The tank crews represented three companies, with each company comprised of three platoons of five tank crews each. The performance of the 45 tank crews was scored on two occasions, once when they carried out the mission in daytime and once at night. A single observer scored the performance of each crew according to the detailed ARTEP guidelines. Scores for the sample ranged from a low of 210 to a high of 1150. These scores will be referred to as Table VIII data.

Design of the G Study

Admittedly, the designers of the Table VIII data collection did not have generalizability theory in mind. As a consequence, the universe of generalization resolved itself into observation occasions which were confounded with days, time of day, and observer. The full design, taking into account the hierarchical structure in which crews were nested, is a Company (*C*) x Platoon:Company (*P:C*) x Crew:Platoon:Company (*Cr:P:C*) x Occasion (*O*) partially nested design. In words, crews (5) were nested within platoons, and platoons (3) were nested within each company (3); performance was measured on two different occasions (day and night).

Classical Reliability Approach

In classical theory, the best estimate of reliability is the correlation between tank crew scores obtained on two occasions. For a performance score

averaged over the two occasions and ignoring the effect of platoon and company, the reliability is .64.

Clearly, this reliability coefficient is influenced by leniency of different observers, the difficulty of the terrain or terrains on which the missions were conducted, the differences between missions, the time of day (day or night), the day that the performance was observed, and so forth. However, the importance of these possible sources of measurement error cannot be estimated using classical theory, even if the measurement facets had been systematically identified. Furthermore, performance might be influenced by the policies and leadership skills within particular companies or platoons. Classical reliability is mute on how to treat these hierarchical data.

Generalizability Theory Approach

The generalizability analysis proceeded along the lines suggested by symmetry:

(1) Choose the facets of measurement and compute mean squares.
(2) Estimate variance components.
(3) Specify the facet (or combination of facets) that is the focus of measurement, and specify the sources of error.
(4) Examine alternative D-study designs.

Steps 1 and 2 are shown in Table 15 for the Company (*C*) x Platoon: Company (*P:C*) x Crew: Platoon: Company (*Cr:P:C*) x Occasion (*O*) partially nested design.

Interpretation of Variance Components In theory, a variance component cannot be negative, yet a negative estimate occurred (as indicated in

TABLE 15 Variance Components for the Study of Tank-Crew Performance Measurement[a]

Source of Variation	Mean Squares	Estimated Variance Component
Companies (*C*)	55461	0[b]
Platoons:C (*P:C*)	78636	1607.19
Crews:P:C (*Cr:P:C*)	45383	15967.50
Occasions (*O*)	244505	3573.21
Cx	83711	3538.79
P.C x *O*	30629	3436.17
Cr:P:C x *O*	31448	13448.20

[a] The design is crews nested in platoons nested in companies crossed with occasions.
[b] Negative variance component set to 0.

Table 15). With sample Table VIII data, a negative variance component can arise either due to sampling error or misspecification of the measurement model. If the former, the most widely accepted practice is to set the variance coefficient to 0, as was done in Table 15. If the latter, the model should be respecified and variance components estimated with the new model. The rationale for setting the company variance component to 0 was the following. First, the difference in the mean performance of the three companies was small: 770.90, 763.33, and 692.93. Variation among company means accounted for only 0.3 percent of the total variation in the data. The best estimate of the variance due to companies, then, was 0. (See the concluding section for additional discussion on estimating variance components.)

The largest variance component in Table 15 is for crews: the universe-score variance. Crew performance differs systematically, and the measurement procedure reflects this variation. The next largest component is associated with the residual, indicating that error is introduced due to inconsistency in tank-crew performance from one occasion to the next, and other unidentified sources of error (e.g., inconsistency due to time of day, observer, terrain, and the like). The remaining variance components are roughly one-fourth the size of the residual, with the exception of the component for companies. Since the variance component for companies is 0 and the variance component for platoons is the smallest one remaining, neither sufficiently influences variation in performance enough to have an important influence if they are considered part of the universe-score variance.

Generalizability Coefficients. Since decision makers are interested in the generalizability of unit performance, one possible method for calculating the G coefficient for crews is:

$$\hat{\rho}^2_{crews} = \frac{\hat{\sigma}^2_{CrP:C}}{\hat{\sigma}^2_{CrP:C} + \hat{\sigma}^2_{o} / n'_o + \hat{\sigma}^2_{res} / n'_o} = 0.65$$

The generalizability of tank crew performance, averaged over the two observation occasions, is .65. If, however, the decision maker is interested in the generalizability of the score of a single tank crew selected randomly and observed on a single occasion, the coefficient drops to .48 due to the large residual variance component.

The principle of symmetry states that the universe-score variance is comprised of all components that give rise to systematic variation among crews. In this case, variation due to companies and platoons, as well as variation due to crews, must be considered universe-score variation. Characteristics of companies and platoons, such as leadership ability, contribute to systematic variation among crews. Following symmetry, the G coefficient for crews, averaged over two occasions, is:

$$\hat{\rho}^2_{crews*} = \frac{\hat{\sigma}^2_{CrP:C} + \hat{\sigma}^2_{P:C} + \hat{\sigma}^2_{C}}{\hat{\sigma}^2_{CrP:C} + \hat{\sigma}^2_{P:C} + \hat{\sigma}^2_{C} + \hat{\sigma}^2_{O}/n'_o + \hat{\sigma}^2_{P:CO}/n'_o + \hat{\sigma}^2_{CO}/n'_o + \hat{\sigma}^2_{res}/n'_o} = 0.59$$

We write $\rho^{2crews*}$ to distinguish this coefficient from the one above.

Surprisingly, by increasing universe-score variance, the G coefficient decreased, for two reasons. The increase in universe-score variance by incorporating systematic variation due to companies was negligible:

$$\hat{\sigma}^2_{C} = 0,\ \hat{\sigma}^2_{P:C} = 1607.10$$

And the additional error introduced ($\sigma^2_{p.co}$ and σ^2_{co}) by considering variation due to companies and platoons as universe-score variance, while not large relative to other sources of variation (e.g., σ^2_{res}), were large relative to the systematic variability of companies and platoons.

Finally, if the decision maker is interested in the dependability of platoon performance, the generalizability of the measurement was estimated (aggregating over crews within platoons and occasions) as follows:

$$\hat{\rho}^2_{Platoons} = \frac{\hat{\sigma}^2_{P:C}}{\hat{\sigma}^2_{P:C} + \hat{\sigma}^2_{CrP:C}/n'_{CrP:C} + \hat{\sigma}^2_{O}/n'_o + \hat{\sigma}^2_{P:CO}/n'_o + \hat{\sigma}^2_{res}/n'_{CrP:C}n'_o} = 0.17$$

Notice here that *crews* is considered a source of error; variability in crews introduces uncertainty in estimating the performance of the entire platoon—the average of the performance of a platoon's individual crews. The low generalizability coefficient, then, reflects the fact that there was greater variability among crews within a platoon than among platoons.

CONCLUDING COMMENTS ON GENERALIZABILITY THEORY: ISSUES AND LIMITATIONS

In the preceding sections, I argued that generalizability theory was the most appropriate behavioral measurement theory for treating military performance measures and showed how the theory could be used to model and improve performance measures. Even the best of theories have limitations in their applications, and generalizability theory is no exception. In concluding, I address the following topics: negative estimated variance components; assumption

of constant universe scores; and dichotomous data (for a more extensive treatment, see Shavelson and Webb, 1981; Shavelson et al., 1985).

Small Samples and Negative Estimated Variance Components

Two major contributions of generalizability theory are its emphasis on multiple sources of measurement error and its deemphasis of the role played by summary reliability or generalizability coefficients. Estimated variance components are the basis for indexing the relative contribution of each source of error and the *un*dependability of a measurement. Yet Cronbach et al. (1972) warned that variance-component estimates are unstable with usual sample sizes of, for example, a couple of occasions and observers. While variance-component estimation poses a problem for G theory, it also afflicts *all* sampling theories. One virtue of G theory is that it brings estimation problems to the fore and puts them up for examination.

Small Samples and Variability of Estimated Variance Components

The problem of fallible estimates can be illustrated by expressing an expected mean square as a sum of population variances. In a two-facet, crossed (*p* x *i* x *j*), random model design, the variance of the estimated variance component for persons—of the estimated universe-score variance—is

$$VAR(\sigma_p^2) = \frac{2}{n_{p-1}}\left[\left(\sigma_p^2 + \frac{\sigma_{pj}^2}{n_j} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{res}^2}{n_j n_i}\right)^2 + \frac{1}{(n_j - 1)}\left(\frac{\sigma_{pj}^2}{n_j} + \frac{\sigma_{res}^2}{n_j n_i}\right)^2 + \frac{1}{n_i - 1}\left(\frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{res}^2}{n_j n_i}\right)^2 + \frac{1}{(n_j - 1)(n_i - 1)}\left(\frac{\sigma_{res}^2}{n_j n_i}\right)^2\right]$$

With all of the components entering the variance of the estimated universe-score variance, the fallibility of such an estimate is quite apparent, especially if $n(i)$ and $n(j)$ are quite modest. In contrast, the variance of the estimated residual variance has only one variance component,

$$VAR(\sigma_{res}^2) = \frac{2}{\left[(n_p - 1)(n_i - 1)(n_j - 1)\right]}\sigma_{res}^4 .$$

In a crossed design, then, the number of components and hence the variance of the estimator increase from the highest-order interaction component to the main effect components. Consequently, sample estimates of the universe-score variance—estimates of crucial importance to the dependability of a measurement—may reasonably be expected to be less stable than estimates of components of error variance.

Negative Estimates of Variance Components

Negative estimates of variance components can arise because of sampling errors or because of model misspecification (Hill, 1970: see also previous discussion). With respect to sampling error, the one-way ANOVA illustrates how negative estimates can arise. The expected mean squares are:

$$E\ \mathrm{MS}_{Within} = \sigma^2_1,$$

and

$$E\ \mathrm{MS}_{Between} = \sigma^2_1 + n\sigma^2_2 = \sigma^2_{1\ 2},$$

where $E\ \mathrm{MS}_{\mathrm{Within}}$ is the expected value of the mean square within groups and $E\ \mathrm{MS}_{\mathrm{Between}}$ is the expected value of the mean square between groups. Estimation of the variance components is accomplished by equating the observed mean squares with their expected values and solving the linear equations. If $\mathrm{MS}_{\mathrm{Within}}$ is larger than $\mathrm{MS}_{\mathrm{Between}}$ the estimate of σ^2_2 will be negative.

Realizing this problem in G theory, Cronbach et al. (1972:57) suggested that a plausible solution is to substitute zero for the negative estimate, and carry this zero forward as the estimate of the component when it enters any equation higher in the table of mean squares.

Notice that by setting negative estimates to 0, the researcher is implicitly saying that a reduced model provides an adequate representation of the data, thereby admitting that the original model was misspecified. Although solutions such as Cronbach et al.'s are reasonable, the sampling distribution of the (once negative) variance component as well as those variance components whose calculation includes this component is more complicated and the modified estimates are biased. Brennan (e.g., 1983) provides an alternative algorithm that sets all negative variance components to 0. Each variance component, then, "is expressed as a function of mean squares and sample sizes, and these do not change when some other estimated variance component is negative" (Brennan, 1983:47). Brennan's procedure produces unbiased estimated-variance components, except for negative components set to 0.

Bayesian methods provide a solution to the problem of negative variance-component estimates (e.g., Box and Tiao, 1973; Davis, 1974; Fyans, 1977; Shavelson and Webb, 1981). Consider a design with two sources of variation: within groups and between groups. The Bayesian approach includes the constraint that MS(between groups) is greater than or equal to MS(within groups) so that the between-groups variance component cannot be negative. Unfortunately, the computational complexities involved and the distributional-form assumptions make these procedures all but inaccessible to practitioners.

An attractive alternative that produces nonnegative estimates of variance components is maximum likelihood (Dempster et al., 1981). Maximum likelihood estimators are functions of every sufficient statistic and are consistent and asymptotically normal and efficient (Harville, 1977). Although these estimates are derived under the assumption of a normal distribution, estimators so derived may be suitable even with an unspecified distribution (Harville, 1977). Maximum likelihood estimates have not been used extensively in practice because they are not readily available in popular statistical packages. However, researchers at the University of California, Los Angeles, (Marcoulides, Shavelson, and Webb) are examining a restricted maximum likelihood approach that, in simulations so far, appears to offer considerable promise in dealing with the negative variance component problem.

Assumption of Constant Universe Scores

Nearly all behavioral measurement theories assume that the behavior being studied remains constant over observations; this is the steady-state assumption made by both classical theory and G theory. Assessment of stability is much more complex when the behavior changes over time. Among those investigating time-dependent phenomena are Bock (1975), Bryk and colleagues (Bryk and Weisberg, 1977; Bryk et al., 1980), Rogosa and colleagues (Rogosa, 1980; Rogosa et al., 1982, 1984).

Rogosa et al. (1984) consider generalizability theory as one method for assessing the stability of behavior over time. Their approach is to formulate two basic questions about stability of behavior: (1) Is the behavior of an individual consistent over time? (2) Are individual differences among individuals consistent over time?

For individual behavior, consistency is defined as absolutely invariant behavior over time. They characterized inconsistency in behavior in several ways: unsystematic scatter around a flat line, a linear trend (with and without unsystematic scatter), and a nonlinear trend (with or without scatter). Changing behavior over time has important implications in generalizability theory for the estimation of universe scores. When behavior changes systematically over time, the universe-score estimate will be time dependent.

The second, and more common, question about stability is the consistency of individual differences among individuals. Perfect consistency occurs whenever the trends for different individuals are parallel, whether the individuals' trends are flat, linear, or nonlinear.

A generalizability analysis with occasions as a facet is described by Rogosa et al. (1984) as one method for assessing the consistency of individual differences over time. The variance component that reflects the stability of individual differences over time is the interaction between individuals and occasions. A small component for the interaction (compared to the variance component for universe scores) suggests that individuals are rank-ordered similarly across occasions; that is, their trends are parallel. It says nothing about whether individual behavior is changing over time. As described above, the behavior of *all* individuals could be changing over time in the same way (a nonzero main effect for occasions). A relatively large value of the component for the individuals x occasion interaction (compared to the universe-score variance component) shows that individuals are ranked differently across occasions. This could be the result of unsystematic fluctuations in individual behavior over time, the usual interpretation made in G theory under the steady-state assumption. But it could also reflect differences in systematic trends over time for different individuals. The behavior of some individuals might systematically improve over time, while that of others might not. Furthermore, the systematic changes could be linear or nonlinear.

Clearly, it is necessary to specify the process by which individual military performance changes in order to model this change. Rogosa et al. provide excellent steps in that direction by describing analytic methods for assessing the consistency of behavior of individuals and the consistency of differences among individuals. At the least, their exposition is valuable for clarifying the limited ability of G theory to distinguish between real changes in behavior over time and random fluctuations over time that should be considered error.

Although the analytic models for investigating time-dependent changes in behavior are important, they do not alleviate the investigator's responsibility to define the appropriate time interval for observation. In studying the dependability of a measurement, it is necessary to restrict the time interval so that the observations of behavior can reasonably be expected to represent the same phenomenon.

There are other developments in the field that examine changing behavior over time, such as models of change based on Markov processes (e.g., Plewis, 1981). However, since these developments do not follow our philosophy of isolating multiple sources of measurement error, and do not provide much information about how measurement error might be characterized or estimated, they are not discussed here.

Dichotomous Data

Analysis of variance approaches to reliability, including G theory, assume that the scores being analyzed represent continuous random variables. When the scores are dichotomous, as they were in the earlier example with observers' "go-no go" scores for soldiers' performance on the revolver task, analysis of variance methods produce inaccurate estimates of variance components and reliability (Cronbach et al., 1972; Brennan, 1980). In analyses of achievement test data with dichotomously scored items, L. Muthén (1983) found that the ANOVA approach for estimating variance components tended to overestimate error components and underestimate reliability. She found that a covariance structure analysis model (see B. Muthén, 1978, 1983; Jöreskog, 1974), specifically designed to treat dichotomous data as a manifestation of an underlying continuum (B. Muthén, 1983), produced estimates of variance components and generalizability coefficients that were closer to the true values than those from the ANOVA.

Concluding Comment

Used wisely, none of the foregoing limitations invalidates G theory. They simply point to the care needed in designing and interpreting the results of G studies.

In spite of its limitations, generalizability theory does what those seeking to determine the dependability of performance measures want a theory of behavioral measurement to do. G theory:

(1) models the sources of error likely to enter into a performance measurement,
(2) models the ways in which these errors are sampled,
(3) provides information on where the major source of measurement error lies,
(4) provides estimates of how the measurement would improve under alternative plans for sampling and thereby controlling sources of error variance, and
(5) indicates when the measurement problem cannot be overcome by sampling, so that alternative revisions of the measurement (e.g., modifications in administration, training of observers, or both) might be considered.

REFERENCES

Bell, J.F. 1985 Generalizability theory: the software problem. *Journal of Educational Statistics* 10:19-30.

Bock. D. 1975 *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

Box, G.E.P., and G.C. Tiao 1973 *Bayesian Inference in Statistical Analysis*. Reading, Mass.: Addison-Wesley.

Brennan. R.L. 1980 Applications of generalizability theory. In R.A. Berk, ed., *Criterion -Referenced Measurement: The State of the Art*. Baltimore. Md.: The Johns Hopkins University Press.

1983 *Elements of Generalizability Theory*. Iowa City, Iowa: American College Testing Publications.

Bryk, A.S., and H.I. Weisberg 1977 Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin* 84:950-962.

Bryk, A.S., J.F. Strenio, and H.I. Weisberg 1980 A method for estimating treatment effects when individuals are growing. *Journal of Educational Statistics* 5:5-34.

Cardinet. J., and L. Allal 1983 Estimation of generalizability parameters. Pp. 17-48 in L.J. Fyans, Jr., ed., *Generalizability Theory: Inferences and Practical Applications* . San Francisco: Jossey-Bass.

Cardinet, J., and Y. Tourneur 1974 The Facets of Differentiation [sic] and Generalization in Test Theory. Paper presented at the 18th congress of the International Association of Applied Psychology, Montreal, July-August.

1977 *Le Calcul de Marges d'Erreurs dans la Theorie de la Generalizabilite* . Neuchatel. Switzerland: Institut Romand de Recherches et de Documentation Pedagogiques.

Cardinet, J.Y. Tourneur, and L. Allal 1976a The generalizability of surveys of educational outcomes. Pp. 185-198 in D.N.M. DeGruijter and L.J. Th. van der Kamp, eds., *Advances in Psychological and Educational Measurement*. New York: Wiley.

1976b The symmetry of generalizability theory: applications to educational measurement. *Journal of Educational Measurement* 13:119-135.

1981 Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement* 18:183-204.

Cronbach, L.J. 1976 Research on Classrooms and Schools: Formulation of Questions, Design, and Analysis. Occasional paper, Stanford Evaluation Consortium. Stanford University (July).

Cronbach, L.J., G.C. Gleser, A.N. Nanda, and N. Rajaratnam 1972 *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Davis, C.E. 1974 Bayesian Inference in Two-way Analysis of Variance Models: An Approach to Generalizability. Unpublished doctoral dissertation. University of Iowa.

Dempster, A.P., D.B. Rubin, and R.K. Tsutakawa 1981 Estimation in covariance components models. *Journal of the American Statistical Association* 76:341-353.

Erlich, O., and R.J. Shavelson 1976 The Application of Generalizability Theory to the Study of Teaching. Technical Report 76-9-1, Beginning Teacher Evaluation Study. Far West Laboratory, San Francisco.

Fyans, L.J. 1977 A New Multi-Level Approach for Cross-Cultural Psychological Research. Unpublished doctoral dissertation. University of Illinois at Urbana-Champagne.

Hartley, H.O.. J.N.K. Rao. and L. LaMotte 1978 A simple synthesis-based method of variance component estimation. *Biometrics* 34:233-242.

Harville, D.A. 1977 Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72:320-340.

Hill, B.M. 1970 Some contrasts between Bayesian and classical influence in the analysis of variance and in the testing of models. Pp. 29-36 in D.L. Meyer and R.O. Collier, Jr.. eds.. *Bayesian Statistics*. Itasca, Ill.: F.E. Peacock.

Jöreskog, K.G. 1974 Analyzing psychological data by structural analysis of covariance matrices. In D.H. Krantz, R.C. Atkinson, R.D. Luce, and P. Suppes, eds., *Contemporary Developments in Mathematical Psychology*, Vol. II. San Francisco: W.H. Freeman & Company.

Kahan, J.P., N.M. Webb, R.J. Shavelson, and R.M. Stolzenberg 1985 *Individual Characteristics and Unit Performance: A Review of Research and Methods*. R-3194-MIL. Santa Monica, Calif.: The Rand Corporation.

Muthén, B. 1978 Contributions to factor analysis of dichotomous variables. *Psychometrika* 43:551-560.

1983 Latent variable structural equation modeling with categorical data. *Journal of Econometrics* 22:43-65.

Muthén, L. 1983 The Estimation of Variance Components for the Dichotomous Dependent Variables: Applications to Test Theory. Unpublished doctoral dissertation, University of California, Los Angeles.

Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics) 1983 Second Annual Report to the Congress on Joint-Service Efforts to Link Standards for Enlistment to On-the-Job Performance. A report to the House Committee on Appropriations. U.S. Department of Defense, Washington, D.C.

Plewis, I. 1981 Using longitudinal data to model teachers' ratings of classroom behavior as a dynamic process. *Journal of Education Statistics* 6:237-255.

Rogosa, D. 1980 Comparisons of some procedures for analyzing longitudinal panel data. *Journal of Economics and Business* 32:136-151.

Rogosa, D., D. Brandt, and M. Zimowski 1982 A growth curve approach to the measurement of change. *Psychological Bulletin* 90:726-748.

Rogosa, D., R. Floden, and J.B. Willett 1984 Assessing the stability of teacher behavior. *Journal of Educational Psychology* 76:1000-1027.

Rubin, D.B., reviewer 1974 The dependability of behavioral measurements: theory of generalizability for scores and profiles. *Journal of the American Statistical Association* 69:1050.

Shavelson, R.J. 1985 *Evaluation of Nonformal Education Programs: The Applicability and Utility of the Criterion-Sampling Approach*. Oxford, England: Pergamon Press.

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

Shavelson, R.J., N.M. Webb, and, L. Burstein 1985 The measurement of teaching. In M.C. Wittrock. ed., *Handbook of Research on Teaching*, 3rd ed. New York: Macmillan.

Tourneur, Y. 1978 Les Objectifs du Domaine Cognitif. 2me Partie: Theorie des Tests. Ministere de l'Education Nationale et de la Culture Francaise, Universite de l'Etat a Mons, Faculte des Sciences Psycho-Pedagogiques.

Tourneur, Y., and J. Cardinet 1979 Analyse de Variance et Theorie de la Generalizabilite: Guide pour la Realization des Calculs. Doc. 790.803/CT/9. Universite de l'Etat a Mons, France.

U.S. Department of Labor 1972 *Handbook for Analyzing Jobs*. Washington, D.C.: U.S. Department of Labor.

Webb, N.M., and R.J. Shavelson 1981 Multivariate generalizability of general educational development ratings. *Journal of Educational Measurement* 18:13-22.

Webb, N.M., R.J. Shavelson, J. Shea, and E. Morello 1981 Generalizability of general educational development ratings of jobs in the U.S. *Journal of Applied Psychology* 66:186-191.

Webb. N.M., R.J. Shavelson, and E. Maddahian 1983 Multivariate generalizability theory. Pp. 67-82 in L. J. Fyans, Jr., ed., *Generalizability Theory: Inferences and Practical Applications* . San Francisco: Jossey-Bass.

Wittman, W.W. 1985 Multivariate reliability theory: principles of symmetry and successful validation strategies. Pp. 1-104 in R.B. Cattell and J.R. Nesselroade, eds., *Handbook of Multivariate Experimental Psychology*, 2nd ed. New York : Plenum Press.

Procedures for Eliciting and Using Judgments of the Value of Observed Behaviors on Military Job Performance Tests

Richard M. Jaeger and Sallie Keller-McNulty

THE PROBLEMS ADDRESSED

As part of a Joint-Service job performance measurement project, each Service is developing a series of standardized hands-on job performance tests. These tests are intended to measure the "manifest, observable job behaviors" (Committee on the Performance of Military Personnel, 1984:5) of first-term enlistees in selected military occupational specialties. Once the tests have been constructed and refined, they will be examined for use as criteria for validating the Armed Services Vocational Aptitude Battery (ASVAB), or its successor instruments, as devices for classifying military enlistees into various service schools and military occupational specialties.

Three problems are addressed in this paper. The first concerns the development of standards of minimally acceptable performance on the newly developed criterion tests. Such standards could be used to discriminate between enlistees who would not be expected to exhibit satisfactory (or, perhaps, cost-beneficial) on-the-job performance in a military occupational specialty and those who would be expected to exhibit such performance.

The second problem concerns methods for eliciting and characterizing judgments on the relative value or worth of enlistees' test performances that are judged to be above the minima deemed necessary for admission to one or more military occupational specialties. Practical interest in this problem derives from the need to classify enlistees into military occupational specialties

in a way that maximizes their value to the Service while satisfying the enlistees' own requirements and interests.

The third problem concerns the use of enlistees' behaviors on the hands-on tests, and judgments of their value, in the classification of enlistees among military occupational specialties. As was true of the second problem, interest in this problem reflects the need to assign enlistees to military occupational specialties in a way that satisfies the needs of the Services and the enlistees. In a scarce-resource environment, it is essential that the classification problem be solved in a way that maximizes the value of available personnel to the Services while maintaining the attractiveness of the Services at a level that will not diminish the pool of available enlistees.

The three problems considered in this paper are not treated at the same level of detail. Since there is an extensive methodological and empirical literature on judgmental procedures for setting standards on tests, we have addressed this topic in considerable detail. There is little research that supports methodological recommendations on assigning relative value or worth to various levels of test performance. Therefore, our treatment of this problem is comparatively brief. Finally, our discussion of the problem of assigning enlistees to the military occupational specialties should be viewed as illustrative rather than definitive. This problem is logically related to the first two, but is of such complexity that complete development is beyond the scope of this paper.

Establishing Test Standards

To fulfill the requirements of a military occupational specialty, an enlistee must be capable of performing dozens, if not hundreds, of discrete and diverse tasks. Indeed, each Service has conducted extensive analyses of the task requirements of each of its jobs (Morsch et al., 1961; Goody, 1976; Raimsey-Klee, 1981; Burtch et al., 1982; U. S. Army Research Institute for the Behavioral and Social Sciences, 1984) that have produced convincing evidence of the complexity of the various military occupational specialties and the need to describe military occupational specialties in terms of disjoint clusters of tasks. Even when attention is restricted to the job proficiencies expected of personnel at the initial level of skill defined for a military occupational specialty, the military occupational specialty might be defined by several hundred tasks that can reasonably be allocated to anywhere from 2 to 25 or more disjoint clusters (U.S. Army Research Institute for the Behavioral and Social Sciences, 1984:12-19).

In view of the complexity of military occupational specialties, it is unlikely that the performance of an enlistee on the tasks that compose a military occupational specialty could validly be characterized by a single test

score. In their initial development of performance tests, the service branches have acknowledged this reality by (1) defining clusters of military occupational specialty tasks; (2) identifying samples of tasks that purportedly represent the population of tasks that compose a military occupational specialty; and (3) specifying sets of measurable behaviors that can be used to assess enlistees' proficiencies in performing the sampled tasks. The problem of defining minimally acceptable performance in a military occupational specialty must therefore be addressed by defining minimally acceptable performance on each of the clusters of tasks that compose the military occupational specialty. Methods for defining standards of performance on task clusters thus provide one major focus of this paper.

Eliciting and Combining Judgments of the Worth of Job Performance Test Behaviors

Scores on the job performance tests that are currently under development are to be used as criterion values in the development of algorithms for assigning new enlistees to various military occupational specialties. Were it possible to develop singular, equivalently scaled, equivalently valued measures that characterized the performance of an enlistee in each military occupational specialty, optimal classification of enlistees among military occupational specialties would be a theoretically simple problem. In reality, the problem is complicated by several factors. First, as discussed above, the tasks that compose a military occupational specialty are not unidimensional. Second, even tests that assessed enlistees' performances on task clusters with perfect precision and validity would not be inherently equivalent. Third, the worth or value associated with an equivalent level of performance on tests that assessed proficiency in two different task clusters would likely differ across those clusters. Fourth, the worth or value associated with a given proficiency level in a single task cluster would likely differ, depending on the military occupational specialty in which the task cluster was imbedded.

To address these issues, the problem of establishing functions and eliciting judgments that assign value to levels of proficiency in various military occupational specialties (hereafter called "value functions") must be examined at the level of the individual tasks and at the level of the task clusters. In this regard, two of the major problems considered in this paper are equivalent.

To develop value functions for military occupational specialties, several component problems must be addressed. First, the task clusters defined by job analysts for each military occupational specialty must be accepted or revised. Second, value functions associated with performances on tasks sampled from task clusters must be defined. Third, operational procedures

for eliciting judgments of the values of various levels of performance on tasks sampled from task clusters must be developed. Fourth, methods for weighting and aggregating value assignments across sampled tasks, so as to determine a value assignment for a profile of performances on the tasks that are sampled from a military occupational specialty, must be developed. Related issues that must be considered include the comparability of value assignments across tasks within a military occupational specialty, as well as the scale equivalence of value assignments to levels of performance in different military occupational specialties.

Using Predicted Test Performances and Value Judgments in Personnel Classification

Assuming it is possible to predict enlistees' performances on military job performance tests from the ASVAB or other predictor batteries, and assuming that judgments of the values of these predicted performances can be elicited and combined to produce summary scores for military occupational specialties, there remains the problem of using these summaries in classifying enlistees among military occupational specialties. This problem can be addressed in several ways, depending on one's desire to consider as primary the interests of individual enlistees and/or the Services, and the types of decision scenarios envisioned.

If it was desired to satisfy the interests of individual enlistees with little regard for the needs of or costs to the military, predicted performances in various military occupational specialties would be used solely for guidance purposes. The only value functions that would be pertinent would be those of the individual enlistee. Enlistees would be assigned to the military occupational specialties they most desired, after having been informed of their likely chances of success in each.

If the interests of the military were viewed as primary, the best classification strategy would depend on the decision scenarios envisioned and the decision components to be taken into account. In a scenario in which each enlistee was to be classified individually, based on his/her predicted military occupational specialty job performances and the set of available military jobs at the time of his/her classification, the obvious classification choice would be the one that carried maximum value. In a scenario in which enlistees were to be classified as a group (e.g., the group of enlistees who completed the ASVAB during a given week), the predicted job performances of all members of the group, and the values associated with those predictions could be taken into account, in addition to the average values associated with the performances of personnel currently assigned to military occupational specialties with jobs available at the time of classification.

These alternatives are considered in a discussion of the problem of using

enlistees' predicted scores on job performance tests in classifying enlistees among military occupational specialties. A specific mathematical programming model for the third alternative is developed and illustrated.

ESTABLISHING MINIMUM STANDARDS OF PERFORMANCE

One of the two major problems considered in this paper is the establishment of standards of performance that define minimally acceptable levels of response on the new criterion tests that are under development by the Services. In addressing this problem, we first discuss the consequential issues associated with standard setting. We next describe the most widely used standard-setting methods that have been proposed for use with educational achievement tests. In the third section, we consider the prospects for applying these methods to the problem of setting standards on military job performance tests. Finally, we examine a variety of operational questions that arise in the application of any standard-setting procedure, such as the types and numbers of persons from whom judgments on appropriate standards are sought, the form in which judgments are sought, and the information provided to those from whom judgments are sought. Rather than recommending the one "best" standard-setting procedure, it is our intent to illuminate the alternatives that have been applied elsewhere, to bring forth the principal considerations that affect their applicability in the military setting, and to bring to light the major operational issues that must be addressed in using any practical standard-setting procedure.

Consequences of Setting Standards

There are no objective procedures for setting test standards. It is necessary to rely on human judgment. Since judgments are fallible, it is important to consider the consequences of setting standards that are unnecessarily high or low. If an unnecessarily high standard is established, examinees whose competence is acceptable will be failed. Errors of this kind are termed false-negative errors. If the standard established is lower than necessary, examinees whose competence is unacceptable will be passed. Errors of this kind are termed false-positive errors. Both individuals and society are placed at risk by these kinds of errors.

When tests are used for selection—that is, for determining who is admitted to an educational program or an employment situation—society or institutions bear the primary effects of false-positive errors. The effects of false-negative errors are borne primarily by individuals when applicant pools greatly exceed institutional needs. However, limitations in the pool of personnel available for military service increase the institutional consequences of making false-negative errors. Adequate military staffing depends on the availability of personnel for a variety of military occupational specialties.

Since the military now relies on an all-volunteer force, it is particularly vulnerable to erroneous exclusion of qualified personnel.

When tests are used for purposes of classification—that is, for allocating examinees among alternative educational programs or jobs—the effects of false-positive *and* false-negative errors are shared by institutions and individuals. When false-positive errors are made, individuals are assigned to programs or jobs that are beyond their levels of competence. This results in less-than-optimal utilization of personnel and the possibility of costly damage for institutions. It also results in psychological and physical hazards for individuals. When false-negative errors are made, individuals are not assigned to programs or jobs for which they are competent. Although this is unlikely to result in physical damage to individuals or institutions, it does produce less-than-optimal use of personnel by institutions and the risk of psychological distress for individuals.

In the military context, the risk to human life and the national security associated with false-positive classification errors is particularly great. Although they might cause psychological distress, false-negative classification errors are unlikely to be life-threatening for individuals. But the Services compete with the civilian sector for qualified personnel. Therefore, the military consequences of false-negative classification errors are likely to be severe for military occupational specialties that require personnel with rare skills and abilities.

Conventional Standard-Setting Procedures

The number of procedures that have been proposed for setting standards on pencil-and-paper tests has been estimated as somewhere between 18 (Hambleton and Eignor, 1980) and 31 (Berk, 1985). The difference between these figures has more to do with the authors' criteria for identifying methods as "different" than with substantively new developments during the years 1980 to 1985. These same authors, as well as others (Meskauskas, 1976; Berk, 1980; Hambleton, 1980), have proposed a variety of schemes for classifying standard-setting procedures. Since this review of standard-setting procedures will be restricted to those that have been widely used and/or hold promise for use in establishing standards on military job performance tests, a simple, two-category classification method will be used. Procedures that require judgements about test items will be described apart from procedures that require judgments about the competence of examinees.

Procedures That Require Judgments About Test Items

Many of the procedures used for setting standards on achievement tests are based on judgments about the characteristics of dichotomously scored

tests items and examinees' likely performances on those items. Both the types of judgments required and the methods through which judgments are elicited differ across procedures. The most widely used procedures of this type are reviewed in this section.

The Nedelsky Procedure. This standard-setting procedure is, perhaps, of historical interest since it is the oldest procedure in the modern literature on standard setting that still enjoys widespread use. It was proposed by Nedelsky in 1954, and is only applicable to tests composed of multiple-choice items.

The first step in the procedure is to define a population of judges and to select a representative sample from that population. Judges who use the procedure must conceptualize a "minimally competent examinee" and then predict the behavior of this minimally competent examinee on each option of each multiple-choice test item. Because of the nature of the judgment task, it is essential that judges be knowledgeable about the proficiencies of the examinee population, the requirements of the job for which examinees are being selected, and the difficulties of the test items being judged.

For each item on the test, each judge is asked to predict the number of response options a minimally competent examinee could eliminate as being clearly incorrect. A statistic termed by Nedelsky the "minimum pass level" (MPL) is then computed for each item. The MPL for an item is equal to the reciprocal of the number of response options remaining, following elimination of the options that could be identified as incorrect by a minimally competent examinee. The test standard based on the predictions of a single judge is computed as the sum of the MPL values produced by that judge for all items on the test.

An initial test standard is computed by averaging the summed MPL values produced by the predictions of each of a sample of judges. Nedelsky (1954) recommended that this initial test standard be adjusted to control the probability that an examinee whose true performance was just equal to the initial test standard could be classified as incompetent due solely to measurement error in the testing process. The adjustment procedure recommended by Nedelsky depends on the assumption that the standard deviation of the test standards derived from the predictions of a sample of judges is equal to the standard error of measurement of the test. If the assumption were correct, and if the distribution of measurement errors on the test were normal, the probability of failing an examinee with true ability just equal to the initial recommended test standard could be reduced to any desired value. For example, reducing the initial test standard by one standard deviation of the distribution of summed MPL values would ensure that no more than 16 percent of examinees with true ability just equal to the initial recommended test standard would fail. Reducing the initial recommended test standard by two standard deviations would reduce this probability to about 2 percent.

The initial recommended test standard produced by Nedelsky's procedure derives from the assumption that examinees will make random choices among the item options that they cannot eliminate as being clearly incorrect. Examinees are assumed to have no partial information or to be uninfluenced by partial information when making their choices among remaining options. If these assumptions were correct, and if judges were able to correctly predict the average number of options a minimally competent examinee could eliminate as being clearly incorrect, the initial tests standard resulting from the Nedelsky procedure would be an unbiased estimate of the mean tests score that would be earned by minimally competent examinees. However, studies by Poggio et al. (1981) report that, when Nedelsky's procedure was applied to pencil-and-paper achievement tests in a public school setting, school personnel were unable to make consistent judgments of the type required to satisfy the assumptions of the procedure.

The Angoff Procedure. Although he attributes the procedure to Ledyard Tucker (Livingston and Zieky, 1983), William Angoff's name is associated with a standard-setting method that he described in 1971. The procedure requires that each of a sample of judges consider each item on a test and estimate (1971:515):

> the *probability* that the "minimally acceptable" person would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score.

As was true of Nedelsky's procedure, the first step in using Angoff's procedure is to identify an appropriate population of judges and then to select a representative sample from this population. Judges are then asked to conceptualize a minimally competent examinee. Livingston and Zieky (1982) suggest that judges be helped to define minimal competence by having them review the domain that the test is to assess and then take part in a discussion on what constitutes "borderline knowledge and skills." If judges can agree on a level of performance that distinguishes between examinees who are competent and those who are not, Zieky and Livingston recommend that the definition of that performance be recorded, together with examples of performance that are judged to be above, and below, the standard. Using as an example a test that was designed to assess the reading comprehension of high school students, Zieky and Livingston suggest that judges be asked to reach agreement on whether a minimally competent student must be able to "find specific information in a newspaper article, distinguish statements of fact from statements of opinion, recognize the main idea of a paragraph," and so on. To be useful in characterizing a

minimally competent examinee, the behaviors used to distinguish between those who are competent and those who are not should represent the domain of behavior assessed by the test for which a standard is desired.

The judgments required by Angoff's procedure are as follows: Each judge, working independently, considers the items on a test individually and predicts for each item the probability that a minimally competent examinee would be able to answer the test item correctly.

The sum of the probabilities predicted by a judge becomes that judge's recommended test standard and, if the predictions were correct, would equal the total score on the examination that would be earned by a minimally competent examinee. The average of the recommended test standards produced by the entire sample of judges is the test standard that results from Angoff's procedure.

If for each item on the test the average of the probabilities predicted by the sample of judges was correct, the test standard produced by Angoff's procedure would equal the mean score earned by a population of minimally competent examinees. In any case, the result of Angoff's procedure can be viewed as a subjective estimate of that mean.

Angoff's procedure has been modified in several ways, so as to make it easier to use and/or to increase the reliability of its results. One modification involves use of a fixed scale of probability values from which judges select their predictions. This technique allows judges' predictions to be processed by an optical mark-sense reader for direct entry to a computer, this saving a coding step and reducing the possibility of clerical errors. Educational Testing Service used an asymmetric scale of probabilities when setting standards on the subtests of the National Teacher Examinations (NTE). Livingston and Zieky (1982:25) objected to the use of an asymmetric scale, since they felt it might bias judges' predictions. Cross et al. (1984) used a symmetric scale of 10 probability values that covered the full range from zero to one, thus overcoming Livingston and Zieky's objections.

Other modifications of Angoff's procedure include the use of iterative processes through which judges are given an opportunity to discuss their initial predictions and then to reconsider those predictions. Cross et al. (1984) investigated the effects of such a process coupled with the use of normative data on examinees' actual test performances. They found that judges recommended a lower test standard at the end of a second judgment session than at the end of an initial session. These results were not entirely consistent with findings of Jaeger and Busch (1984) in a study of standards set for the National Teacher Examinations. They found that mean recommended standards were lower at the end of a second judgment session than at the end of an initial session for four out of eight subtests of the NTE Core Battery; they found just the reverse for the other four subtests. However, the variability of recommended test standards was consistently reduced by

using an iterative judgment process. The resulting increase in the stability of mean recommended test standards suggests that use of an iterative judgment process with Angoff's procedure is advantageous.

The Ebel Procedure. The Ebel (1972:492-494) standard-setting procedure also begins by defining a population of judges and selecting a representative sample from that population. After conceptualizing a "minimally competent" examinee, judges must complete three tasks.

First, judges must construct a two-dimensional taxonomy of the items in a test, one dimension being defined by the "difficulty" of the test items and the other being defined by the "relevance" of the items. Ebel suggested using three levels of difficulty, which he labeled "easy," "medium," and "hard." He suggested that four levels of item relevance be labeled "essential," "important," "acceptable," and "questionable." However, the procedure does not depend on the use of these specific categories or labels. The numbers of dimensions and categories could be changed without altering the basic method.

The judges' second task is to allocate each of the items on the test to one of the cells created by the two-dimensional taxonomy constructed in the first step. For example, Item 1 might be judged to be of "medium difficulty" and to be "important;" Item 2 might be judged to be of "easy difficulty" and to be of "questionable" relevance, etc.

The judges' final task is to answer the following question for each category of test items (Livingston and Zieky, 1982:25):

> If a borderline test-taker had to answer a large number of questions like these, what percentage would he or she answer correctly?

When a test standard is computed using Ebel's method, a judge's recommended percentage for a cell of the taxonomy is multiplied by the number of test items the judge allocated to that cell. These products are then summed across all cells of the taxonomy to produce a recommended test standard for that judge. As in the procedures described earlier, the recommendations of all sampled judges are averaged to produce a final recommended test standard.

The Jaeger Procedure. This procedure was developed for use in setting a standard on a high school competency test (Jaeger, 1978, 1982), but can be adapted to any testing situation where a licensing, certification, or selection decision is based on an examinee's test performance (Cross et al., 1984).

One or more populations of judges must be specified, and representative samples must be selected from each population. As in the procedures described above, judges are asked to render judgments about test items. More specifically, judges are asked to answer the following question for each item

on the test for which a standard is desired: Should *every* examinee in the population of those who receive favorable action on the decision that underlies use of the test (e.g., every enlistee who is admitted to the military occupational specialty) be able to answer the test item correctly? Notice that this question does not require judges to conceptualize a "minimally competent" examinee.

An initial standard for a judge is computed by counting the number of items for which that judge responded "yes" to the question stated above. An initial test standard is established by computing the median of the standards recommended by each sampled judge.

Jaeger's procedure is iterative by design. Judges are afforded several opportunities to reconsider their initial recommendations in light of data on the actual test performances of examinees and the recommendations of their fellow judges. In its original application, judges were first asked to provide "yes/no" recommendations on each test item on a 120-item reading comprehension test. The judges were then given data on the proportion of examinees who had actually answered each test item correctly in the most recent administration of the test, in addition to the distribution of test standards recommended by their fellow judges. Following a review of these data, judges were asked to reconsider their initial recommendations and once again answer, for each item, the question of whether every "successful" examinee should be able to answer the test item correctly. These answers were used to compute a new set of recommended test standards in preparation for a final judgment session. Prior to the final judgment session, judges were given data on the proportion of examinees who completed the test during the most recent administration who would have failed the test had the standard been set at each of the score values between zero and the maximum possible score. In addition, judges were shown the distribution of test standards recommended by their fellow judges during the second judgment session. A final judgment session, identical to the first two, was then conducted. The "yes" responses were tabulated for each judge, and the final recommended test standard was defined as the median of the standards computed for each judge.

Jaeger (1982) recommends that more than one population of judges be sampled, and that the final test standard be based on the lowest of the median recommended standards computed for the various samples of judges. He also suggests that prior to the initial judgment session each judge complete the test under conditions that approximate those used in an actual test administration.

Procedures That Require Judgments About Examinees

Unlike the standard-setting procedures that have been described to this point, several widely used procedures do not require judgments about the

characteristics or difficulty of test items. Instead, judges are asked to make decisions regarding the competence of individual examinees on the ability measured by the test for which a standard is sought. Proponents of these procedures claim that the types of judgments required—concerning persons rather than test items—are more consistent with the experience and capabilities of educators and supervisory personnel. The resulting test standards are thus claimed to be more reasonable and realistic.

The Borderline-Group Procedure. This standard-setting procedure was proposed by Zieky and Livingston (1977). As is true of all standard-setting procedures, the first step in applying the procedure is to define an appropriate population of judges and then to select a representative sample from that population. Livingston and Zieky (1982:31) indicate the importance of sampled judges knowing, or being able to determine, the level of knowledge or skill in the domain assessed by the test of individual examinees they will be asked to judge. Careful and appropriate selection of judges is thus critical to the success of the procedure.

Judges are first asked to define three categories of competence in the domain of knowledge or skill assessed by the test: "adequate or competent," "borderline or marginal," and "inadequate or incompetent." Ideally, these definitions would be operational, would be consensual, and would be reached collectively following extensive deliberation and discussion by the entire sample of judges. In reality, this ideal might not be achieved, nor might a process of face-to-face discussion among judges be feasible.

Once definitions of the three categories of competence have been formulated, the principal act of judgment in the borderline-group procedure requires judges to identify members of the examinee population whom they would classify as "borderline or marginal" in the knowledge and/or skill assessed by the test. It is essential that the judges use information other than the score on the test for which a standard is sought in reaching their classification decisions. If scores on this test were used, the standard-setting procedure would be tautological. Additionally, classification decisions based on scores on the test for which a standard is sought might well be biased. Interpretations of test performances are often normative, and individual judges are unlikely to know about the performances of a representative sample of the population of examinees.

The test for which a standard is sought is administered, under standardized conditions, after a subpopulation of examinees has been classified as "marginal or borderline." The standard produced by the borderline-group method is defined as the median of the distribution of test scores of examinees who are classified as "marginal or borderline."

Although the borderline-group procedure has some definite advantages,

it is subject to several factors that threaten the validity of the test standard it produces. First, unless the sample of examinees that is classified by the judges is in its distribution of test scores representative of the population of examinees to which the test standard is applied, a biased standard will result. Second, in making their classifications it is essential that judges restrict their attention to knowledge and/or skill that is assessed by the test for which a standard is sought. To make reasoned decisions, judges must be familiar with the performance of examinees they are to classify. However, the better they know these examinees, the more likely they are to be influenced in their judgments by factors other than the knowledge and/or skill assessed by the test; halo effect is a pervasive influence in judgments that require classification of persons. Finally, the "borderline" category is the middle position of a three-point scale that ranges from "competent" to "incompetent." Numerous studies of judges' classification behavior have shown that the middle category of a rating scale tends to be used when judges do not have information that is sufficient to make a valid judgment. Contamination of the "borderline" category with examinees that do not belong there would bias the test standard produced by the borderline-group procedure.

The Contrasting-Groups Procedure. Proposed by Zieky and Livingston in 1977, this procedure is similar in concept to the criterion-groups procedure suggested by Berk (1976). The principal focus of judgment in the contrasting-groups procedure is on the competence of examinees rather than the characteristics of test items, just as in the borderline-group procedure.

The first two steps of the contrasting-groups procedure are identical to those of the borderline-group procedure. First, a population of judges is defined and a representative sample of judges is selected from that population. Second, the sampled judges must develop operational definitions of three categories of competence in the domain of knowledge and/or skill assessed by the test for which a standard is sought: "adequate or competent," "borderline or marginal," and inadequate or incompetent."

The principal judgmental act of the contrasting-groups procedure requires judges to assign a representative sample of the population to be examined to the three categories of competence they have just defined. That is, each member of the sample of examinees is assigned to a category labeled "adequate or competent," "borderline or marginal, or "inadequate or incompetent."

Once classification of examinees has been completed, the test for which a standard is sought is administered to the examinees about whom judgments have been made. The standard that results from the contrasting-groups method is based on the test score distributions of examinees who

have been assigned to the "adequate or competent" and "inadequate or incompetent" categories.

Several methods have been proposed for analyzing the test score distributions of examinees who have been assigned to the "adequate or competent" and the "inadequate and incompetent" categories. Hambleton and Eignor (1980) recommended that the two test score distributions be plotted on the same graph and that the test standard be defined as the score at which these two distributions intersect. This procedure assumes that the score distributions will not be coincident and that they will be overlapping. Under these conditions, the test standard that results from this algorithm has the advantage of minimizing the total number of examinees who were classified as "competent" and who would fail the test, plus the total number of examinees who were classified as "incompetent" and who would pass the test. If the loss attendant to passing an incompetent examinee were not equal to the loss attendant to failing a competent one, this test standard would not minimize total expected losses. However, if the loss ratio was either known or estimable, the standard could be adjusted readily so as to minimize expected losses.

Livingston and Zieky (1982) proposed an alternative method of analyzing the test score distribution of "competent" examinees for the purpose of setting a standard. They suggested that the percentage of examinees classified as "competent" be computed for the subsample of examinees who earned every possible test score. The test standard would be defined as the test score for which 50 percent of the examinees were classified as "competent." Since for small samples of examinees, the distribution of test scores is likely to be irregular, Livingston and Zieky (1982) recommend the use of a smoothing procedure prior to computing the score value for which 50 percent of the examinees were classified as "competent." They describe several alternative smoothing procedures.

Most of the cautions enumerated above for the borderline-group procedure apply to the contrasting-groups procedure as well: Judges must have an adequate and appropriate basis for classifying examinees, yet avoid classification on bases outside the domain of knowledge and/or skill assessed by the test. A representative sample of examinees must be classified so as to avoid distortion of the distributions of test scores of "competent" and "incompetent" examinees. Since not only the shapes of test score distributions but the sample sizes on which they are based will affect their point of intersection, use of a representative sample of examinees is essential to the fidelity of the standard resulting from the contrasting-groups procedure.

Berk's (1976) criterion-groups procedure is operationally identical to the contrasting-groups procedure apart from the definition of groups. In Berk's method, instead of classifying examinees as "competent" or "incompetent,"

"criterion" groups are formed from examinees who are "uninstructed" and "instructed" in the material assessed by the test for which a standard is sought. Of course, judgment is needed to define groups that can appropriately be termed "uninstructed" and "instructed." A fundamental assumption of Berk's method is that the "uninstructed" group is incompetent in the knowledge and/or skill assessed by the test, and that the "instructed" group is competent in that knowledge and/or skill.

Prospects for Applying Conventional Standard-Setting Procedures

Although three Services are developing new pencil-and-paper tests as components of their job performance criterion measures (Laabs, 1984; Committee on the Performance of Military Personnel, 1984), a principal interest of the military is establishment of standards on the performance components of the new measures. Since all of the conventional standard-setting procedures reviewed above were developed for use with pencil-and-paper tests in a public education setting, they might not be applicable to hands-on and/or interview procedures used in a military setting. We will consider the applicability of the procedures in the order of their initial description.

Procedures That Require Judgments About Test Items

The Nedelsky procedure may be only partially applicable in setting standards on military job performance tests because it can be used only with multiple-choice test items, while the assessment of "manifest, observable job behaviors" is a central purpose of the military job performance tests. The performance components of these tests (in the Joint-Service lexicon, the hands-on portions) typically measure active performance of a task in accordance with the specifications of a military manual. Because the behavior to be measured is appropriate action, not discrimination among proposed actions, a multiple-choice item format would appear to be inconsistent with specified measurement objectives.

The Nedelsky procedure could be used to establish standards on the "knowledge measures" portion of the criterion measures being developed by the Army, and on similar tests developed by other Services, provided the measures consist of items in multiple-choice format. In civilian settings, the Nedelsky procedure often has provided standards that are somewhat more lenient than those provided by other procedures. In the proposed military setting, lenient standards of performance on individual tasks still lead to stringent standards on an entire job performance test. This would be true if a separate standard had to be established for each task and satisfactory performance were required on all sampled tasks. For example, suppose that pencil-and-paper measures had been developed for 10 tasks that composed

a military occupational specialty job performance test. If the standard of performance adopted for each measure resulted in just 5 percent of enlistees failing, and if examinee performances on the various measures were independent (an admittedly unlikely occurrence, used here merely to illustrate the extreme case), the percentage of examinees who would satisfy the overall military occupational specialty criterion on the pencil-and-paper portion of the job performance test would be 100 x (1-0.05)10 = 59.9 percent. Thus almost 40 percent of the examinees would fail the pencil-and-paper portion of the job performance test, even though only 5 percent would fail to complete any given task. An alternative standard-setting procedure that resulted in more stringent standards for each task would result in an even higher (and perhaps unacceptable) failure rate on the pencil-and-paper portion of the job performance test.

The stimulus question that defines the fundamental judgment task of the Nedelsky procedure could be stated in any of several seemingly reasonable ways. A central issue would be the appropriate referent for a "minimally competent examinee." An example using the Army military occupational specialty (MOS) 95B (military police) should clarify the issue. Suppose that one tested task from this military occupational specialty was "restraining a suspect." Should the judges being asked to recommend a standard for the test of knowledge of this task be asked to:

> Think about a soldier who has just been admitted to MOS 95B who is borderline in his/her knowledge of restraining a subject. Which options of each of the following test items should this soldier be able to eliminate as obviously incorrect?

Or should the judges be asked to:

> Think about a soldier who has just been admitted to MOS 95B who is just borderline in his/her knowledge needed to function satisfactorily in that MOS. Which options of each of the following test items should this soldier be able to eliminate as obviously incorrect?"

The difference here is in the referent population. One is task-specific and the other refers to the entire military occupational specialty. Either choice could be supported through logical argument. Since the test is task-specific, the task-delimited population is consistent with Nedelsky's specifications. However, the tested task is one of many that could have been sampled from those that compose the military occupational specialty and the domain of generalization of fundamental interest is the military occupational specialty rather than the sampled task. Perhaps the stimulus question should be constructed on practical rather than purely logical grounds. Judges could be asked whether it is easier to conceptualize a minimally competent soldier who has just been admitted to the military occupational specialty or a soldier who was minimally competent in performing the task being tested.

Some experiments could be conducted to determine the referent population that produced the smallest variation in recommended test standards. Mean recommended standards resulting from use of the two referent populations could be compared to determine whether they differed and which appeared to be most reasonable.

The standard-setting procedures proposed by Angoff, Ebel, and Jaeger can be used with any test that is composed of items or activities that are scored dichotomously. Since all of the military job performance measures we have reviewed are of this form, any of these standard-setting procedures could be used with these tests.

Like the Nedelsky procedure, both the Angoff and Ebel procedures require that judges define a minimally competent examinee. The issue of an appropriate referent population, discussed in the context of the Nedelsky procedure, would therefore be of concern with these procedures as well. Once the question of an appropriate referent population was settled, adaptation of the Angoff procedure to military job performance tests with dichotomously scored components would appear to be straightforward. For example, when used with the performance component of a criterion test for military occupational specialty 95B, the stimulus question might be:

> Think about 100 soldiers who have just been admitted to MOS 95B who are borderline in their ability to restrain a suspect. What percentage of these 100 soldiers would position the suspect correctly when applying handcuffs?

Similar questions could be formed for each tested activity in the "restraining a suspect" task.

Ebel's procedure might not be applicable to military job performance tests for several practical reasons. The procedure presumes that the "items" that compose a test are unidimensional but stratifiable on dimensions of difficulty and relevance. Many of the military job performance measures we have reviewed contain very few activities or items, so that stratification of items might not be possible. Asking a judge "What percentage of the items on this test should a minimally competent examinee be able to answer correctly?" is tantamount to asking "Where should the test standard be set?". Without stratification of items into relatively homogeneous clusters, Ebel's method is unlikely to yield stable results. Theoretically, Ebel's method could be applied to an overall job performance test to yield a standard for an entire military occupational specialty rather than a single task. However, several assumptions inherent in the method would then be highly questionable. The most obvious basis for stratification of items or activities on the test would be by task, but it is not likely that the activities or items used to assess an examinee's performance of a single task are homogeneous in relevance or difficulty. Also, the relative lengths of subtests that assess

an examinee's performance of different tasks are probably a consequence of the level of detail contained in the military procedures manual that describes the tasks, rather than the relative importance of the tasks. Since Ebel's method weights item strata in proportion to their size, a task that contained a larger number of activities would receive more weight in determining an overall test standard than would a task that contained fewer activities, regardless of the relative importance of the two tasks. In computing an overall test standard, Ebel's procedure has no provision for weighting item strata by importance or by any other judgmental consideration.

From a purely mechanical standpoint, Jaeger's standard-setting procedure could be used readily with either the pencil-and-paper or performance components of military job performance tests. Since it does not require judges to conceptualize a minimally competent examinee, the problem of defining an appropriate referent population, a central issue with the other item-based standard-setting procedures, would not arise with Jaeger's procedure. However, if the empirical results observed in civilian public-school settings are also found in the military context, Jaeger's procedure might yield test standards that are unacceptably high. If standards are established for tests of each sampled task, this problem is likely to be greatly exacerbated. The principal advantage of Nedelsky's procedure, illustrated above, might well be the principal disadvantage of Jaeger's procedure, since stringent test standards for each task would translate to an impossibly stringent standard for admission to a military occupational specialty.

When using Jaeger's procedure, judges might be asked the following question for each activity in a test designed to assess performance on a designated task: "Should every enlistee who is accepted for this MOS be able to perform this activity?". On the tests we have reviewed, it appears that the activities listed closely mirror descriptions of standard practice as specified in applicable military procedures manuals. If judges based their recommendations on "the book" they would likely answer questions about most, if not all, activities affirmatively, thus resulting in impossibly high test standards.

Our expectation then, is that Jaeger's procedure could be adapted to military job performance tests quite readily, but would likely yield test standards that were impractically high. This expectation should not preclude small-scale empirical investigations of the procedure.

Procedures That Require Judgments About Examinees

When considered for setting standards on military job performance tests, both the borderline-group procedure and the contrasting-groups procedure offer appealing characteristics and features. These procedures can be used with tests composed of any type of test item, regardless of item scoring, as

long as the tests assess some unidimensional variable and yield a total score. Although the small sample of military job performance tests we have reviewed contained tasks that were made up of discrete units that could be treated as "items," some performance tests might not be assembled in this way, thus rendering the item-based standard-setting procedures inapplicable. For example, a test concerned with operation of a simple weapon might be scored on the basis of time to effect firing or accuracy of results. If the variable representing "success" can be scored continuously, the borderline-group or contrasting-groups standard-setting procedures could be used to determine a standard of performance, even though the item-based procedures could not.

A second advantage of the standard-setting procedures that require judgments about examinees is that the types of judgments required are probably similar to those made routinely by military supervisors, both in training schools and active units. In fact, somewhat similar judgments were requested of military job experts in the study published by the U.S. Army Research Institute for the Behavioral and Social Sciences (1984). Appendix A of the Army report contains a scale for assessing the abilities of soldiers to perform various tasks associated with specified military occupational specialties.

Despite their advantages, the borderline-group and contrasting-groups standard-setting procedures present several operational problems that might be difficult to overcome. Both procedures require classification of examinees into groups labeled unacceptable, borderline, and acceptable, and subsequent testing of persons in at least one of these groups. When discussing the item-based standard-setting procedures, we suggested that the appropriate referent population of "minimally competent examinees" was not obvious. In a somewhat different form, the same problem must be dealt with for the person-based procedures: Should judges be asked to classify examinees as "unacceptable," "borderline," or "acceptable" in the skills defined by the task cluster represented by the test or in all skills needed to function within a military occupational specialty? Since a standard is likely to be desired for a test that is restricted to a single task cluster, one could argue that the appropriate referent population is obvious. On the other hand, eliminating personnel who cannot meet all the demands of a military occupational specialty is the ultimate goal.

A second, and perhaps more serious problem, is obtaining judgments, and test data on examinees who are "unacceptable." Under current military classification procedures, such persons would rarely be assigned to active duty in a military occupational specialty. First, potentially unacceptable enlistees are screened out on the basis of their ASVAB performances. Few such persons are assigned to service schools that provide the training necessary to enter a military occupational specialty for which they are "unacceptable."

Second, since success in an appropriate service school is prerequisite to assignment to active duty in a military occupational specialty and screening takes place prior to graduation from military service schools, the population of potentially unacceptable personnel assigned to active duty in a military occupational specialty is further reduced. The contrasting-groups procedure requires the identification of personnel who are "acceptable" and "unacceptable" in the skills assessed by the test for which a standard is sought. The number of "unacceptable" examinees must be sufficiently large to obtain a stable distribution of test scores. If very few "unacceptable" persons are admitted to a military occupational specialty, obtaining a sufficient number of nominations might not be possible. Recall that classification of examinees to the "unacceptable," "borderline," and "acceptable" groups must be based on information other than scores on the tests for which standards are sought. In the present context, that information would have to consist of observations of on-the-job performance of enlistees early in their initial tours of duty in a military occupational specialty. Again, to the extent that current military classification systems are effective, the number of "unacceptable" enlistees will be very small.

Operational Questions and Issues

Regardless of the procedure used to establish standards on military job performance tests, a set of common operational issues must be considered. Since judgments are required, one or more appropriate populations of judges must be identified. The numbers of judges to be sampled from each population must be specified. The stimulus materials used to elicit judgments must be developed. The substance and process of training judges must be specified and developed. The information to be provided judges, both prior to and during the judgment process, must be specified. A decision must be reached on handling measurement error within the process of computing a standard, and if measurement error is to be considered an algorithm for doing so must be developed. We will discuss each of these issues briefly.

Types of Judges to be Used

All of the item-based standard-setting procedures (with the possible exception of Jaeger's procedure) require judges to be knowledgeable about the distribution of ability of examinees on the skills assessed by the test and about the distributions of performance of examinees on each item contained in the test. Judges used with these procedures should therefore have experience in observing and working with the examinee population, either in a service-school setting or, preferably, in the actual working environment of

the military occupational specialty for which the test is a criterion. Judges used with the person-based standard-setting procedures must meet even more stringent criteria. Since they must classify individual examinees, they must be knowledgeable about the abilities of these individuals to perform specific tasks in the actual work settings of the military occupational specialty.

These requirements suggest that either instructional personnel in appropriate military service schools or immediate unit supervisors (such as noncommissioned officers) in military occupational specialty field units could serve as judges for the item-based standard-setting procedures, but only the latter personnel would be suitable as judges for the examinee-based standard-setting procedures.

Numbers of Judges to be Used

In any standard-setting procedure, the numbers of judges to be used should be determined by considering the probable magnitude of the standard error of the recommended test standard as a function of sample size. Since in all of the standard-setting procedures described in this paper, the recommendations of individual judges are derived more or less independently and are aggregated only at the point of computing a final test standard, the standard error of that recommendation will vary inversely as the square root of the size of the sample of judges.

Ideally, the size of the sample of judges would be sufficient to reduce the standard error of the recommended test standard to less than half a raw score point on the test for which the standard was desired. In that case, assuming that the recommended test standard varied normally across samples of judges, the probability that an examinee whose test score was equal to the test standard recommended by a population of judges would pass or fail the test, just due to sampling of judges, would be no more than 0.05.

In practice, the size of a sample of judges needed to reduce the standard error of the recommended test standard to the ideal point might be prohibitively expensive or otherwise infeasible to obtain. An alternative criterion for sample size might be based on the relative magnitudes of the standard error of the recommended test standard and the standard error of measurement of the test for which a standard was desired. Since these sources of error are independent and therefore additive, it is possible to determine the contribution of sampling error to the overall error in establishing a test standard. For example, if the standard error of the mean test standard was half the magnitude of the standard error of measurement of the test, the variance error of the mean would be only one-fourth the variance error of measurement, and the overall standard error would be increased by a factor of (1.25)1/2 = 1.12 or 12 percent. Alternatively, if the standard error of the mean test standard was one-tenth the magnitude of the standard error of

measurement of the test, the variance error of the mean would be only one one-hundredth the variance error of measurement, and the overall standard error would be increased by a factor of $(1.01)^{1/2} = 1.005$, or 0.5 percent.

Empirical work by Cross et al. (1984) and Jaeger and Busch (1984) showed that, for the subtests of the National Teacher Examinations, relative magnitudes of standard errors of the mean and measurement closer to the latter example than the former were realized with samples of 20 to 30 judges. A modified, iterative form of Angoff's standard-setting procedure was used in each of these studies.

Stimulus Materials to be Used in Setting Standards

The specific stimulus materials to be used in a standard-setting procedure must, of course, be based on the steps involved in conducting that procedure. However, it is essential that the materials be explicit and standardized. All judges must engage in the same standard-setting process, must be fully informed about the types of judgments required of them, and must be privy to the same types of information given all other judges. Experience has shown that judges should be given written as well as standardized oral instructions on the purposes for which their judgments are sought, the types of judgments they are to make, and the exact procedures they are to follow.

The questions judges are asked to answer must be developed with caution and care. For example, in the Angoff standard-setting procedure, judges are asked to estimate the probability that a "minimally competent" examinee would be able to answer each test item correctly. Different responses should be expected, depending on whether judges are asked whether examinees "would be able to answer test items correctly" or "should be able to answer test items correctly." The first question requires a prediction of actual examinee behavior, while the second one requires a statement of desired examinee behavior. Another subtle, but important, distinction in the Angoff procedure concerns the issue of guessing. Since the Angoff procedure is frequently applied to multiple-choice items, examinees might well answer a test item correctly by guessing. Judges will likely estimate different probabilities of actual examinee behavior, depending on whether they are told to consider guessing, to ignore the possibility of guessing, or are given no instructions about guessing. The latter procedure is the least desirable, since consideration of guessing on the part of examinees becomes a source of error variance in the responses of judges.

This example illustrates one of many details that must be addressed if the stimulus materials used in a standard-setting procedure are to function correctly. The goal in developing stimulus materials should be to minimize the variance of recommendations across judges due to factors other than true differences in their judgment of an appropriate test standard.

Training of Judges

Obviously, if judges are to provide considered and thoughtful recommendations on standards for military job performance tests, they must understand their tasks clearly and completely. Judges will have to be trained to do their jobs if these ends are to be achieved.

Although the specifics of a training program for judges must depend on the standard-setting procedure used, some common elements can be identified. First, judges must thoroughly understand the test for which standards are desired. One effective way to meet this need is to have judges complete the test themselves under conditions that approximate an operational test administration. This training technique has been used successfully in setting standards on high school competency tests (Jaeger, 1982) and knowledge tests for beginning teachers (Cross et al., 1984; Jaeger and Busch, 1984).

Second, judges must understand the sequence of operations they are to carry out in providing their recommendations. Since in some standard-setting procedures a single set of judgments is elicited, the procedures are likely to be straightforward and easily learned through a single instructional session followed by a period for answering questions. However, other standard-setting procedures are iterative and require judges to provide several sets of recommendations. In these cases, a simulation of the judgment process might be necessary to ensure that judges know what is expected of them. Jaeger and Busch (1984) used such a simulation, together with a small, simulated version of the test for which standards were desired, in a three-stage standard-setting procedure. Following the actual standard-setting exercise, almost all judges reported that they fully understood what they were to do.

Third, in some standard-setting procedures, judges are provided with normative data on the test performances of examinees. Typically, these data are provided in graphical or tabular form, and since the types of graphs or tables used might not be familiar to them, judges might require instruction on their interpretation. For example, in modified versions of the Angoff and Jaeger standard-setting procedures, judges are shown a "p-value" for each item on the test. It should not be assumed that judges will know that these numbers represent the proportions of examinees who answered each test item correctly when the test was last administered. Normative data on examinees' test performances have also been provided in the form of an ogive (cumulative distribution function graph). It is not reasonable to assume that all judges will know how to read and interpret such graphs without specific training.

The overall objective of training should be to ensure that all judges are responding to the same set of questions on the basis of accurate and common understanding of their judgment tasks.

Information to be Provided to Judges

Citing both logical and empirical grounds, several researchers (Glass, 1978; Poggio et al., 1981) have questioned the abilities of judges to make sensible recommendations of the sort required by most item-based standard-setting procedures. In support of their contentions, these authors cite a number of studies in which recommended test standards would have resulted in outlandish examinee failure rates. For example, Educational Testing Service's study to determine standards for the National Teacher Examinations in North Carolina produced recommendations that would have resulted in denial of teacher certification to half the graduates of the state's accredited and approved teacher education programs.

For some military and civilian occupations one could reasonably argue that examinee failure rates are irrelevant to decisions on an appropriate test standard. For example, brain surgeons, jet engine mechanics, and pilots must be able to perform well all tasks that are essential to their jobs. For these types of positions, it is clearly more damaging to employ less than fully competent persons than to have unfilled positions. But for other, less critical jobs, or where on-the-job training might reasonably be used to compensate for marginal qualification at the time an applicant is hired, one could argue that judges' recommendations for test standards should be based on a realistic assessment of the capabilities of the examinees to whom the test will be administered, as well as the requirements of the job itself.

In such situations, it has been recommended that judges be provided with normative information on examinees' test performances to enable them to evaluate the consequences of their proposed test standards. As mentioned in the preceding section, several iterative standard-setting procedures provide judges with item p-values as well as cumulative distributions of the total scores earned by a representative sample of examinees during the most recent administration of the test. Studies have shown that judges use such information when formulating their recommendations (Jaeger, 1982) and that the predominant effect of such data is to reduce the variability of judges' recommendations (Cross et al., 1984; Jaeger and Busch, 1984).

The principal logical argument in support of such use of normative test data is that judges who are well informed on the capabilities of the examinees to whom the test will be administered are likely to provide more reasoned (and therefore better) recommendations on appropriate test standards. That normative test data also appear to reduce the variability of judges' recommended test standards, thereby increasing the reliability of their recommendations, is a serendipitous finding that adds nothing to the logical argument.

Another type of information judges might be provided during an iterative standard-setting procedure is a summary of the recommendations of their

fellow judges. A large body of social psychological literature, dating from the work of Sherif (1947), suggests that most persons are influenced by the judgments of their peers in decision situations. The manner in which information on the judgments of peers is provided has a crucial influence on the outcome of the judgment process. Summary data in the absence of justification might induce a shift in judgment toward the central tendency of the group, thereby reducing variability, but are unlikely to result in better informed, and hence more reasoned, judgments. A more defensible procedure would allow judges to state and justify the reasons for their recommendations. If this procedure is followed, it is essential that it be carefully controlled to avoid domination by one or a few judges, and to ensure that a full spectrum of judgments is explained.

Measurement Error

Errors of measurement on tests typically are assumed to be normally distributed (Gulliksen, 1950; Lord and Novick, 1968). Based on this assumption, a person whose true level of ability was equal to the standard established for a test would have a 50 percent chance of earning an observed score that fell below the standard, just due to errors of measurement. A person whose true level of ability was one standard error of measurement above the standard would still have a 16 percent chance of earning an observed score that fell below the standard.

To protect against the possibility of failing an examinee as a result of measurement error, several researchers have proposed that initial test standards be adjusted downward by some multiple of the standard error of measurement of the test for which a standard is desired. As described above, Nedelsky (1954) recommended such an adjustment as an integral part of his standard-setting procedure. Unfortunately, he falsely assumed that the standard error of measurement of a test would be well approximated by the standard deviation of judges' recommended standards and based his adjustment on the latter value.

It is unnecessary to adopt Nedelsky's assumption. In most standard-setting situations, the standard error of measurement of the test can be estimated through an internal-consistency reliability estimation procedure, if by no other means, and the recommended test standard can be adjusted accordingly.

Whether a test standard should be adjusted to compensate for errors of measurement is an arguable point, and some might even suggest that the proper adjustment is upward, rather than downward. At issue in the current application are the relative merits of protecting enlistees, or the military, from the consequences of failing a competent examinee or passing an incompetent

one. It is likely that such consequences will vary substantially across tasks and military occupational specialties.

DEFINITION AND CONSTRUCTION OF VALUE FUNCTIONS

The problems discussed in this section concern the establishment of functions that assign value (or worth) to different levels of proficiency in completing various military occupational specialty tasks, and the use of these value functions in assessing the overall worth of an enlistee in a specific military occupational specialty. A method is proposed for defining a value function for a given task. The use of task value functions to establish an overall military occupational specialty value function is then demonstrated. The discussion focuses particularly on hands-on job sample testing.

As in all evaluation processes, judgments must be made. In this situation judgments are needed on the value or worth of task performance levels. Methods for eliciting these judgments are discussed, as are operational issues that are common to several methods.

Defining Task Value Functions

Psychological decision theory (Zedeck and Cascio, 1984) and social behavior theory (Kaplan, 1975) would appear to lend some insights into the problem of establishing task value functions. Psychological decision theory evaluates an individual's (or institution's) decision making strategy by studying the behavior of the individual (or institution). It is in this somewhat circular way that individual (or institutional) behavior is assigned a value. Social behavior theory is not unlike psychological decision theory. In social behavior theory, the value of an individual's behavior is judged on the basis of existing information about that individual. In both of these theoretical frameworks, the behavior being evaluated is the ability to discriminate among proposed actions.

In carrying out their job tasks, enlistees are to behave in accordance with the specifications of a military manual. We assume that the type of behavior (or performance) of most interest in military job performance tests is enlistees' abilities to begin and carry through specific activities. If this is true, judges have less need to judge enlistees' abilities to discriminate among proposed actions than to evaluate their ability to begin a specific task and carry it through to some level of completion. The value assigned will likely depend on an enlistee's level of completion of the task and how accurately the enlistee adhered to the specifications that define the particular task.

One way to define task value functions would be to treat the problem as a multiattribute-utility measurement problem (Gardiner and Edwards, 1975).

In this setting, each "dimension" of each task performance would be evaluated separately. The dimensions of a task would be defined as the set of measurable behaviors (hereafter called a "performance set") being used to assess enlistees' proficiencies in performing the task. Based on the job performance tests we have reviewed, task performance sets might contain such dimensions as knowledge (measured by a pencil-and-paper test), speed (measured by time taken to complete a hands-on test), and fidelity (measured by the total number of successes on sequences of dichotomously scored hands-on test activities). The same performance set would be used in determining minimally acceptable performance levels when establishing job performance test standards. A task value function would be defined as a weighted average of the values assigned to each of the dimensions in the performance set.

Consider an example: Suppose the task to be assessed is an enlistee's proficiency at putting on a field or pressure dressing. Dimensions of this task could be defined as "pressure administered properly" (measured by the total number of successfully completed activities in a "hands-on" job proficiency test) and "time elapsed before pressure administered" (measured by time taken to complete this portion of the job proficiency test). Judgments of the values of varying levels of performance on these two dimensions would have to be elicited. Assuming these judgments could be secured, the value functions for these dimensions might be similar to those in Figure 1. If these were the only dimensions, the value function for this task, $V_{>t}$, would be a weighted average of the value functions for these dimensions. Mathematically,

$$V_t(P) = \begin{cases} \dfrac{w_1 V_1(x) + w_2 V_2(y)}{w_1 + w_2}, & \text{if } V_1(x) > 0 \text{ and } V_2(y) > 0 \\ 0, & \text{otherwise} \end{cases}$$

where P is a set of performance scores on the dimensions in the performance set. In this example the set P contains the time, x, elapsed before pressure was administered and the performance level (or consistency with specified procedures), y, of administering pressure. The weights w_1 and w_2 would be determined by assessing the relative importance of the two dimensions.

Value functions for the dimensions of a task could be constructed to range between 0 and 1. Performance-test score distributions could be used to determine realistic maximum performance levels, which would be assigned values equal to 1. Performance scores below minimally acceptable levels would be assigned values equal to 0. The general location and spread

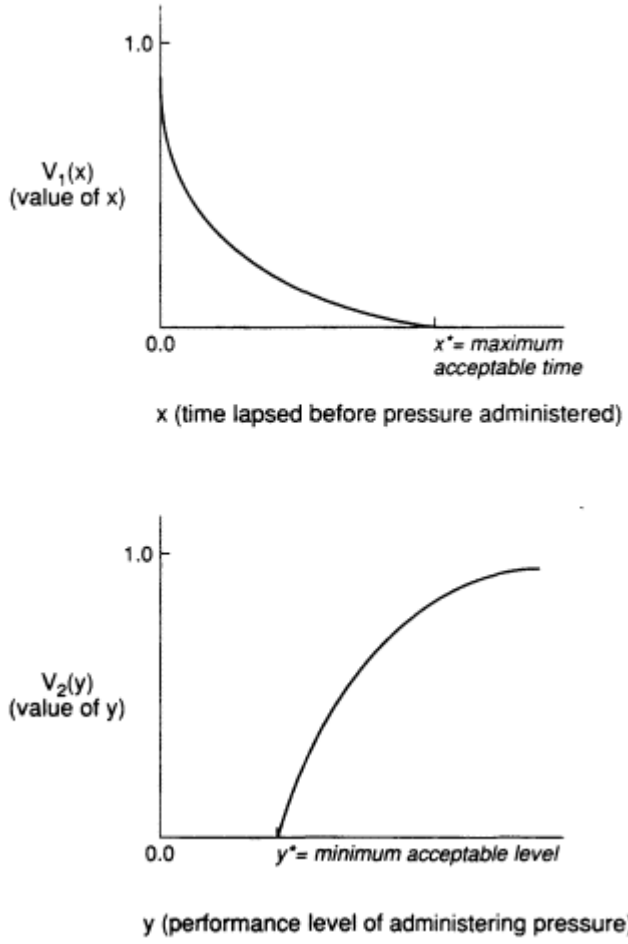


Figure 1 Sample value functions.

of the distributions of performance scores could help define the rates of increase or decrease of value functions.

There is no reason to believe that value functions would be linear. Intuitively, it would seem that small deviations from minimally acceptable performance levels would result in large changes in value, whereas at some higher levels of performance value functions would change more gradually. The actual shapes of value functions would be determined from the judgments elicited on the value (or worth) of different levels of performance (or proficiency) on the dimensions in performance sets.

Structuring Cluster and Military Occupational Specialty Value Functions

A review of the methods used by the Services in choosing the tasks to be included in performance tests indicates that they all used very similar strategies (Morsch et al., 1961; Goody, 1976; Raimsey-Klee, 1981; Burtch et al., 1982; U.S. Army Research Institute for the Behavioral and Social Sciences, 1984). First, each military occupational specialty was partitioned into nonoverlapping clusters, each of which contained similar or related tasks. Judgments were made on the frequency with which each task is performed, on the relative difficulty of each task compared to others in its cluster and the military occupational specialty, and on the relative importance of each task, compared to others in its cluster and the military occupational specialty. Some clusters of tasks appeared in several military occupational specialties and others only appeared in one military occupational specialty. A representative (judgment-based) sample of tasks from each cluster was chosen for the performance test. Based on face validity, it was determined that performance on the tasks sampled from a cluster could be considered to be generalizable to performance on all tasks in the cluster.

If enlistees' performances on the tasks sampled from a cluster are generalizable, the value of their performances on all tasks in that cluster can be computed from a weighted average of the value functions of the individual tasks sampled. The weights should reflect the already determined relative importance of the tasks sampled from the cluster. Note that, with the problem structured in this way, those who make task value judgments need only consider the frequency with which each task is performed and the difficulty of the task. The importance of each task will be reflected in the weights used to compute cluster value functions.

As is true of the task value functions, cluster value functions can be scaled to range between 0 and 1 (inclusive) by dividing the weighted average of task value functions by the sum of the weights, and assigning the cluster value function a value of 0 if performance on any of the sampled tasks in that cluster is below the minimally acceptable level, ($V_t(P) = 0$ for any task t). A cluster value function would have a value close to 1 if performances on all the sampled tasks in that cluster were close to their maximum possible levels. Cluster value functions defined in this way would be on a comparable scale. It would thus be possible to compare an enlistee's value (or worth) over different clusters both within and across military occupational specialties.

Comparability of value functions is essential to the classification of enlistees into the military occupational specialties and to the assignment of duties within a military occupational specialty. If an enlistee had a higher value level for a "first aid" cluster than for a "navigational" cluster, then he/

she could be placed in a military occupational specialty where first aid was more essential than navigation. Once placed in a military occupational specialty, this enlistee could be assigned, if possible, more first aid duties than navigational duties.

It is also possible to define an overall value function for each specific military occupational specialty. Such value functions would be based on the generalizability of enlistees' performances on a sample of tasks within a military occupational specialty to their overall performance in that military occupational specialty. As previously mentioned, the relative importance of all tasks within each military occupational specialty has been determined. A military occupational specialty value function can therefore be defined as a weighted average of the value functions of sampled tasks within the military occupational specialty, where the weights are chosen to reflect the relative importance of the tasks sampled from the military occupational specialty. The military occupational specialty value function can be scaled to range between 0 and 1 (inclusive) by dividing the weighted average of task value functions by the sum of the weights and assigning the military occupational specialty value function to 0 if any task sampled from that military occupational specialty has a value level of 0. The resulting military occupational specialty value functions will then be comparable across military occupational specialties. With military occupational specialty value functions defined in this way, it will be possible to determine the military occupational specialty for which a given enlistee has the greatest value or worth.

The problem of specifying military occupational specialty value functions can be defined symbolically in the following way:

P_i = set of performance scores for individual i;
$V_t(P_i)$ = value function for a specific task, t;
$V_c(Pi)$ = value function for a specific cluster, c, within a military occupational specialty;
$V_{mos}(P_i)$ = value function for a specific military occupational specialty;
W_{tc} = weight for V_t, reflects the relative importance of task t within cluster c;
W_{tmos} = weight for V_t, reflects the relative importance of task t within the military occupational specialty;
n_c = number of tasks sampled from cluster c;
n_{mos} = total number of tasks sampled from the military occupational specialty;
N = total number of individuals currently in the military occupational specialty;

$\overline{V}_{mos}$ = current average value of performances of individuals in the military occupational specialty;

$$V_c(P_i) = \begin{cases} \dfrac{\sum_{t=1}^{n_c} w_{tc} V_t(P_i)}{\sum_{t=1}^{n_c} w_{tc}}, & \text{if } V_t(P_i) > 0 \text{ for all } t; \\ 0, & \text{otherwise} \end{cases}$$

$$V_{mos}(P_i) = \begin{cases} \dfrac{\sum_{t=1}^{n_{mos}} w_{tmos} V_t(P_i)}{\sum_{t=1}^{n_{mos}} w_{tmos}}, & \text{if } V_t(P_i) > 0 \text{ for all } t; \\ 0, & \text{otherwise} \end{cases}$$

and

$$\overline{V}_{mos} = \frac{\sum_{i=1}^{N} V_{mos}(P_i)}{N}.$$

Eliciting Value Judgments

Task value functions must be based on two types of judgments. One type concerns the assignment of value to each possible level of performance on each of the dimensions that compose a task performance set. The other type concerns the relative importance (weights) of the dimensions that compose a performance set. Models for eliciting the second type of judgment exist within the procedures used by the Services to determine the relative importance of tasks that compose a military occupational specialty (Morsch et al., 1961; Burtch et al., 1982; U.S. Army Research Institute for the Behavioral and Social Sciences, 1984). Because these procedures can be adopted in their entirety, they will not be discussed further here.

The first type of judgment is inherently more difficult to elicit, since such judgments involve the assignment of value to continua rather than the more familiar ranking procedure associated with determining a value ordering for objects. Two methods for eliciting the first type of judgment are discussed—an average value function method (Gardiner and Edwards, 1975)

and a method of successive lotteries (Winkler and Hays, 1975). The operational questions that arise in eliciting this type of value judgment are much the same as those that arise in eliciting judgments of appropriate test standards. The two situations are contrasted in this section of the paper.

Average Value Function Procedure

This method was employed by Gardiner and Edwards (1975) in a situation that differs from the present military context. Gardiner and Edwards were evaluating land use regulations (building permits) using a multiattribute-utility measurement framework and considering such dimensions as percent of on-site parking, unit rental, size of development, aesthetics, and density of the proposed development. Even though Gardiner and Edwards were evaluating physical, as opposed to human characteristics, their method might be applicable to the problem of eliciting judgments of the values of performances on military job performance tests.

The average value function method operates as its name implies. Each judge derives his/her own value function for a given dimension and these value functions are then averaged across the judges. Of interest here is the information that is supplied to the judges to assist them in deriving their value functions. The following explanation of the method is set in the context of deriving value functions for dimensions of a military occupational specialty task performance set.

First, judges are told the dimensions in the performance set for which value functions are to be constructed. They are given minimally acceptable performance scores, plausible (not absolute) maximum performance scores, and some other fixed performance scores between the minimally acceptable and plausible maximum values in the performance set. The judges are then given a scenario related to the task in question. For example, suppose the task is to "collect and process evidence." The scenario might be a description of the room or building that needs to be searched and the evidence containers on hand. The judges are told to give a value between 0 and 1 (inclusive) for each of the previously fixed performance levels or simply to draw a graph of their value function, basing their decisions on the value (or worth) of an enlistee's performance in this scenario. The value functions recommended by the judges are then averaged, and this information is given back to the judges.

Other relevant information can also be supplied to the judges at this point in the average value function procedure. This might include the shape, location, and spread of the performance score distributions from which the plausible maximum and minimally acceptable performance scores were determined. The judges are then given a second scenario and asked to repeat their evaluations of the worth of the fixed performance levels, based

on this new scenario. The second scenario must be identical, in terms of task characteristics and difficulty, to the first. However, it must be presented in such a way that this is not apparent to the judges. The final value function for a given dimension in the performance set is the average of the value functions recommended by all judges for that dimension, based on the second scenario.

The average value function procedure has merit, in that the scenarios used can be constructed so as to mirror the scenarios used in "hands-on" performance tests or the scenarios used in assessing the relative importance of tasks within a military occupational specialty (U.S. Army Research Institute for the Behavioral and Social Sciences, 1984). In this way, the judges will consider the same set of circumstances that are imposed on an enlistee when his/her performance is measured. One problem with this method is that the frequency with which each task is performed and the difficulty of the tasks are not directly taken into account.

Method of Successive Lotteries

The method of successive lotteries (also called the method of certainty equivalence), as described by Winkler and Hays (1975), is used to develop utility functions in a decision-theoretic framework. The context of the current problem is not unlike a decision-theoretic framework, in that classification decisions are to be based on an enlistee's value (or worth). Stated in this way, what we have termed "value functions" are the utility functions of a decision-theoretic problem. Consequently, it should be possible to apply the method of successive lotteries to evaluate these value functions.

Consider the problem of constructing a value function for one dimension of one task. For example, let the task be shooting a firearm and let the dimension of interest be accuracy in hitting a stationary target (measured by percent of time on target). The method of successive lotteries would be applied in the following way. First, determine a minimally acceptable performance score (minimum acceptable percent of accuracy) and assign this a value just greater than 0. Then determine the plausible maximum performance score and assign this a value of 1. Select several performance scores between these two levels. The value function is to be evaluated and graphed at these scores. The shape of the value function will be given by the curve that results from connecting the points on this graph.

The evaluation consists of a series of comparisons. For example, suppose that the minimally acceptable accuracy level is 30 percent and the plausible maximum accuracy level is 90 percent. Comparisons between sets of lotteries such as the following are made.

Lottery I: Enlistee A shoots with x percent accuracy ($30 < x < 90$) all of the time.

Lottery II: Enlistee B has a probability of p of shooting with 90 percent accuracy and has a probability of $(1 - p)$ of shooting with 30 percent accuracy.

The judgment that is made is to decide for what value of p one is indifferent with respect to the value (or worth) of enlistees A and B. This indifference point, p, is the value of being x percent accurate (or the magnitude of the value function at performance level x, $V(x)$). If, in the example, x percent is 85 percent, we would expect the indifference point, p, to be close to 1, whereas if x percent is 35 percent we would expect the indifference point to be close to 0. Throughout this evaluation process, Lottery II remains fixed and Lottery I changes only in the sense that the value of x is changed.

With this evaluation method, the frequency with which the task is performed and the difficulty of the task need to be taken into consideration in the determination of indifference points. Since several judges can be involved in this evaluation process, the final value function for a given dimension can be computed as an average (mean or median) of all the judges' value functions (sets of indifference points). To help the judges complete their evaluations, a scenario could also be provided. Analogously to the Jaeger (1982) standard-setting procedure, this evaluation method could be iterated several times, by providing judges with summary information about their fellow judges' initial value functions and information regarding enlistees' actual performance test score distributions before each iteration.

Operational Questions and Issues

With the exception of the issue concerning treatment of measurement error, all of the operational issues and questions that were discussed above (in the section on establishing minimally acceptable standards of performance) are pertinent to the process of eliciting judgments of the value of enlistees' task performances. A separate discussion of operational issues in this section would therefore be largely redundant.

One consideration that is appropriate here but would not be appropriate in the establishment of minimally acceptable performance standards is the use of enlistees themselves to determine value functions. First-term personnel who have been assigned to a military occupational specialty for a reasonable period of time should be capable of judging the value or worth of various levels of performance on tasks that compose that military occupational specialty. Zedeck and Cascio (1984) found that peer ratings of personnel were both reasonable and acceptably reliable.

CLASSIFICATION OF NEW ENLISTEES

All branches of the military have developed computerized personnel allocation systems. These systems have been developed and/or adapted to serve a number of purposes, including: enhancing person-job match (Hendrix et al., 1979; Kroeker and Rafacz, 1983; Roberts and Ward, 1982; Schmitz and McWhite, 1984), lowering attrition rates (Kroeker and Folchi, 1984a), balancing minority representation in certain job classifications and providing equal placement opportunity for minorities in all job classifications (Kroeker and Folchi, 1984b). The material in this section illustrates one application of the ideas discussed earlier in this paper to a computerized personnel allocation system. The illustrations are fictitious and are not necessarily representative of the algorithms in current use in any Service.

Assume a set of performance scores for each potential enlistee can be estimated from his/her scores on an aptitude test such as the ASVAB. Call this set of estimated performance scores $\hat{\boldsymbol{P}}$. Estimates of task value functions, $\hat{V}_t(\hat{\boldsymbol{P}})$, cluster value functions, $\hat{V}_c(\hat{\boldsymbol{P}})$, and various military occupational specialties' value functions, $\hat{V}_{mos}(\hat{\boldsymbol{P}})$, can be found, based on this set of estimated performance scores. Using these estimates, several strategies for classification of enlistees into the different military occupational specialties can be defined. These classification strategies fall into two groups: individual classification strategies and institutional classification strategies.

Individual Classification Strategies

The simplest classification scheme would be to let each enlistee choose his/her preferred military occupational specialty from a pool of military occupational specialties for which his/her predicted performance scores satisfied the minimally acceptable standards. Enlistees' choices would have to be monitored and directed to some degree, so that quotas for the various military occupational specialties would be met. This classification method would not require estimation of any value functions. Consequently, no value judgments of task performance levels would have to be made. Also, this classification method would not require any information about predicted performance scores for enlistees other than the one being classified. A classification decision would be made solely on the basis of predicted performance scores of the enlistee in question and on that enlistee's preferences.

If the interests of the military are given primary consideration and if value functions are properly defined, alternative classification methods that better satisfy military requirements can be derived. The most direct way to use value functions for classification would be to classify each enlistee into the military occupational specialty for which he/she had the highest predicted

$\hat{V}_{mos}(\hat{P})$. Since military occupational specialty quotas would have to be taken into consideration, some enlistees would have to be assigned to military occupational specialties that corresponded to their second-or third-largest $\hat{V}_{mos}(\hat{P})$. This classification method would not require any information other than the individual enlistee's predicted value functions and the military occupational specialty quotas.

An Institutional Classification Strategy

The final classification method proposed in this paper is an institutional strategy as opposed to an individual strategy. The problem of assigning new personnel to military occupational specialties in a way that maximized their value to the military could be formulated in several ways. In keeping with standard operations research terminology, we will call the function that defines the overall military value of a set of personnel classification decisions an "objective function" (Hillier and Lieberman, 1974). One possible goal (from which an objective function could be formed) might be to upgrade the average performance level and the average value (or worth) of individuals assigned to military occupational specialties across all military occupational specialties simultaneously. For $\bar{V}_{mos}$ defined as in Equations 1, it is possible to estimate the current average value of the individuals in each military occupational specialty, $\hat{\bar{V}}_{mos}$. This can be done by taking a random sample of enlistees presently in the military occupational specialty and determining $V>_{\text{mos}}(P_i)$ for each individual, *i*, in the sample and averaging those values. Using the estimates of the current average value of individuals in the military occupational specialties and the new enlistees' predicted military occupational specialty values, it is possible to derive a classification scheme that optimizes the anticipated changes (due to the new enlistees) in the average values of individuals assigned to all of the military occupational specialties. In this classification scheme, individual classification decisions would be based on predicted value functions for the entire group of new enlistees about whom classification decisions are to be made and also on information about enlistees currently assigned to the military occupational specialties.

It is important to understand the advantages to the military of this classification scheme, compared to the value function classification method discussed in the previous section. The optimization invoked by this classification strategy would minimize the decreases, while maximizing the increases, in anticipated average values of individuals assigned to all military occupational specialties. To better see the difference between this classification method and the previous method it will be helpful to consider some fictitious data.

For convenience, assume there are only two military occupational specialties,

MOS1 and MOS2. Assume $\hat{V}_{mos1}$ = .7 and $\hat{V}_{mos2}$= .3. Suppose that, after completing the ASVAB or a similar examination, the predicted value functions for one enlistee are $\hat{V}_{mos1}(\hat{P})$ = .8 and $\hat{V}_{mos2}(\hat{P})$ = .6. Also, suppose that this enlistee has the highest predicted value function for MOS2 among all of the new enlistees who have just completed the aptitude examination. What military occupational specialty assignment for this enlistee would be of maximum benefit to the military?

The value function classification method described in the previous section would place this enlistee in MOS1. This classification method would place him/her in MOS2, and thereby be of maximum benefit to the military. Placing this enlistee in either military occupational specialty would likely help raise the average value of individuals in that military occupational specialty because this enlistee's predicted value levels are higher than the current estimated average values of individuals assigned to both of the military occupational specialties. Since $\hat{V}_{mos1}$ is so much larger than $\hat{V}_{mos2}$, the military's immediate interest would be to assign enlistees to MOS2 who would have the greatest potential to help raise the current average value of individuals already assigned to MOS2 (provided the military's goal is as we stated earlier). Recall that the enlistee under consideration has the highest predicted value function among all new enlistees for whom placement decisions are to be made. Consequently, it is this enlistee who would have the greatest (predicted) ability to help raise the current average value of individuals assigned to MOS2. Had there been other new enlistees with predicted value functions for MOS2 greater than .6, the best classification decision would not have been obvious.

Consider another enlistee from this same example for which $\hat{V}_{mos1}(\hat{P})$ = .65 and $\hat{V}_{mos2}(\hat{P})$ = .25. Where should this enlistee be placed and what effect would he/she have on the average values of individuals assigned to the military occupational specialties? The classification method described in the previous section would have assigned this enlistee in MOS1. Without knowing the predicted value levels of all of the new enlistees and the quotas for MOS1 and MOS2, it is impossible to determine the classification of this enlistee that would minimize the potential negative effect he/she would have on the current average values of the individuals assigned to the military occupational specialties.

A general solution that would achieve the military goal previously described can be determined in the following way. Without loss of generality assume there are only two military occupational specialties, MOS1 and MOS2. Consider forming a two-way table of new enlistees' predicted value functions, as shown in Figure 2. Potential enlistees whose values fell in the (0,0) cell would not be admitted into the Services because their predicted performance scores would fall below the minimally acceptable standards. New enlistees with values falling in the (*0,j*) cells would be assigned to

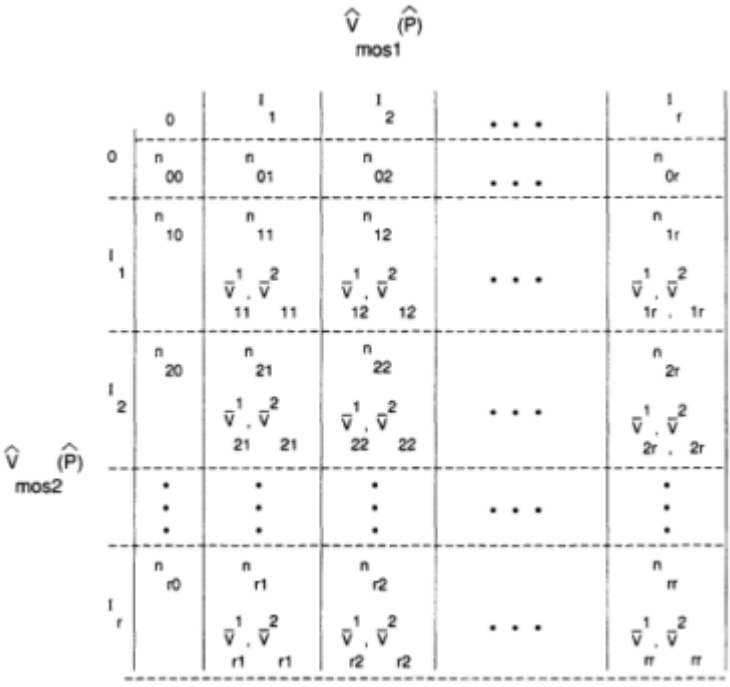


Figure 2 Two-way table of new enlistees' predicted MOS value functions. Notes: I_i = interval of values for $\hat{V}_{mos}(\hat{P})$ n_{ij} = number of new enlistees with $\hat{V}_{mos}(\hat{P})$ in interval I, and $\hat{V}_{mos2}(\hat{P})$ in interval I_j; $\bar{V}_{ij}^{1}$ = average value of $\hat{V}_{mos1}(\hat{P})$ for new enlistees in cell (ij) $\bar{V}_{ij}^{2}$ = average value of $\hat{V}_{mos2}(\hat{P})$ for new enlistees in cell (ij).

MOS1 and those falling in the (i,0) cells would be assigned to MOS2. After these decisions had been made, the quotas could be adjusted to account for the enlistees just assigned to MOS1 and MOS2.

Now, attention can be focused on the remainder of the table. Let Q_1 and Q_2 be the adjusted quotas for MOS1 and MOS2, respectively. Adjust Q_1 and Q_2 such that the number of remaining new enlistees equals the sum of Q_1 and Q_2. For simplicity, assume there are only two intervals of predicted

values, I_1, and I_2. Figure 3 displays the simplified two-way table. Let p_{ij} be the proportion of the new enlistees in the (i,j)th cell to be assigned to MOS1. Let $(1 - p_{ij})$ be the proportion of the new enlistees in the (i,j)th cell to be assigned to MOS2. The predicted average military occupational specialty values for the new enlistees can be expressed as

$$\hat{\bar{V}}^{new}_{mos1} = \left[n_{11} p_{11} \bar{V}^{1}_{11} + n_{12} p_{12} \bar{V}^{1}_{12} + n_{21} p_{21} \bar{V}^{1}_{21} + n_{22} p_{22} \bar{V}^{1}_{22} \right] / Q_1$$

and

$$\hat{\bar{V}}^{new}_{mos2} = \left[n_{11}(1 - p_{11}) \bar{V}^{2}_{11} + n_{12}(1 - p_{12}) \bar{V}^{2}_{12} + n_{21}(1 - p_{21}) \bar{V}^{2}_{21} + n_{22}(1 - p_{22}) \bar{V}^{2}_{22} \right] / Q_2$$

The goal is to find the p_{ij}'s which jointly maximize $\hat{\bar{V}}^{new}_{mos1}$ and $\hat{\bar{V}}^{new}_{mos2}$ while jointly minimizing, if necessary, the amount these values may fall below the current estimated average value of individuals in the military occupational specialties, $\hat{\bar{V}}_{mos1}$ and $\hat{\bar{V}}_{mos2}$.

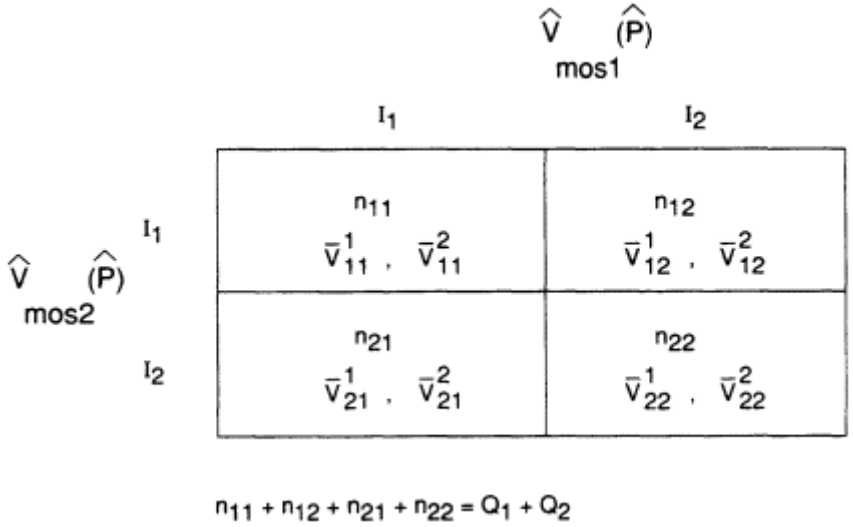


Figure 3 Simplified two-way table of new enlistees' predicted MOS value functions.

This problem can be written mathematically in the following way. Find the p_{ij}'s, Δ_1, and Δ_2 that maximize

$$\left(\hat{\bar{V}}^{new}_{mos1} + \hat{\bar{V}}^{new}_{mos2} - M_1\Delta_1 - M_2\Delta_2\right)$$

subject to

$$\hat{\bar{V}}_{mos1} < \hat{\bar{V}}^{new}_{mos1} + \Delta_1 ;$$

$$\hat{\bar{V}}_{mos2} < \hat{\bar{V}}^{new}_{mos2} + \Delta_2 ;$$

$$0 < \Delta_1 ;$$

$$0 < \Delta_2 ;$$

$$0 < p_{ij} < 1 \text{ for all } (i,j) ;$$

and

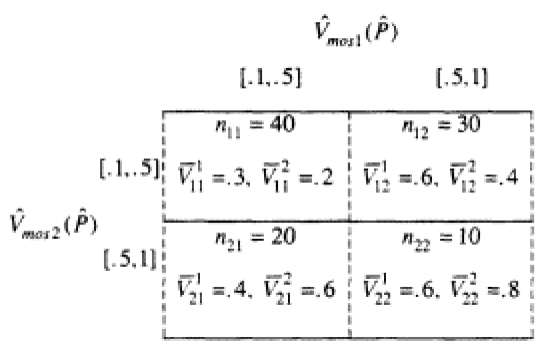
$$n_{11}p_{11} + n_{12}p_{12} + n_{21}p_{21} + n_{22}p_{22} = Q_1 ;$$

where Δ_1 and Δ_2 are the amounts $\hat{\bar{V}}^{new}_{mos1}$ and $\hat{\bar{V}}^{new}_{mos2}$ fall below the current estimated average values of individuals assigned to MOS1 and MOS2, respectively, and M_1 and M_2 are positive known constants. The constants M_1 and M_2 can be thought of as the penalties imposed on the military for admitting enlistees whose predicted performance would result in dropping the average value of individuals in MOS1 and MOS2, respectively. These constants would be chosen by the military.

This formulation of the classification problem is equivalent to a simple linear programming problem that can be solved easily by using the simplex method with the aid of a computer (Hillier and Lieberman, 1974). The formulation can be expanded to include any number of military occupational specialties and any number of value intervals I_i. The following examples have been included to demonstrate the outcome of this classification strategy. The data are fictitious.

Example 1

The following two-way table shows the distribution of 100 new enlistees' predicted value functions for two military occupational specialties.



Let $Q_1 = 55$ and $Q_2 = 45$. Assume estimates of the current average values of personnel currently assigned to the military occupational specialties are $\hat{\bar{V}}_{mos1} = .6$ and $\hat{\bar{V}}_{mos2} = .4$. Let $M_1 = M_2 = 2$. This assigns equal penalties to both military occupational specialties. The linear programming analysis produced the following results:

Cell (*i*,*j*)	P_{ij}	Number of Enlistees Assigned to MOS1	Number of Enlistees Assigned to MOS2
(1,1)	.42	17	23
(1,2)	1.00	30	0
(2,1)	0.00	0	20
(2,2)	.80	8	2
Total		55	45

$\hat{\bar{V}}_{mos1}^{new} = .507$ and $\hat{\bar{V}}_{mos2}^{new} = 401$.

Compare the results of this analysis to those of the following analysis where $M_1 = .5$ and $M_2 = 2$. These choices assign a larger penalty to MOS2 than to MOS1, for decreases in anticipated average values of personnel currently assigned to the military occupational specialties. The linear programming analysis produced the following results:

Cell (*i*,*j*)	P_{ij}	Number of Enlistees Assigned to MOS1	Number of Enlistees Assigned to MOS2
(1,1)	.625	25	15
(1,2)	1.00	30	0
(2,1)	0.00	0	20
(2,2)	0.00	0	10
Total		55	45

$\hat{\bar{V}}_{mos1}^{new} = .464$ and $\hat{\bar{V}}_{mos2}^{new} = .511$.

Example 2

The following two-way table shows the distribution of 180 new enlistees' predicted value functions for two military occupational specialties. This distribution of predicted value functions is similar to that in the example discussed in the text.

$\hat{V}_{mos2}(\hat{P})$ \ $\hat{V}_{mos1}(\hat{P})$	[.1,.3]	(.3,.6]	(.6,1]
[.1,.3]	$n_{11} = 43$ $\bar{V}_{11}^1 = .3,\ \bar{V}_{11}^2 = .2$	$n_{12} = 40$ $\bar{V}_{12}^1 = .6,\ \bar{V}_{12}^2 = .1$	$n_{13} = 25$ $\bar{V}_{13}^1 = .8,\ \bar{V}_{13}^2 = .2$
(.3,.6]	$n_{21} = 26$ $\bar{V}_{21}^1 = .25,\ \bar{V}_{21}^2 = .3$	$n_{22} = 45$ $\bar{V}_{22}^1 = .5,\ \bar{V}_{22}^2 = .33$	$n_{23} = 1$ $\bar{V}_{23}^1 = .8,\ \bar{V}_{23}^2 = .6$
(.6,1]	$n_{31} = 0$	$n_{32} = 0$	$n_{33} = 0$

Let $Q_1 = 90$ and $Q_2 = 90$. Assume estimates of the average values of personnel currently assigned to the military occupational specialties are $\hat{\bar{V}}_{mos1} = .7$ and $\hat{\bar{V}}_{mos2} = .3$. Let $M_1 = 1$ and $M_2 = 3$. These choices assign a larger penalty to MOS2 than to MOS1, for potential decreases in predicted average values. The assignment of penalties in this way is consistent with the military's immediate interest in placing enlistees into MOS2, if they have the greatest predicted potential to help raise the current average value of personnel assigned to MOS2. The linear programming analysis produced the following results:

Cell (i,j)	P_{ij}	Number of Enlistees Assigned to MOS1	Number of Enlistees Assigned to MOS2
(1,1)	.58	25	18
(1.2)	1.00	40	0
(1,3)	1.00	25	0
(2,1)	0.00	0	26
(2,2)	0.00	0	45
(2,3)	0.00	0	1
Total		90	90

$\hat{\bar{V}}_{mos1}^{new} = .572$ and $\hat{\bar{V}}_{mos2}^{new} = .298$.

Notice that, because of the distribution of predicted values of new enlistees it is impossible to raise the average value of personnel assigned to MOS2. However, the optimization process did minimize the decrease in the average value of personnel assigned to MOS2 by allowing the average value of personnel assigned to MOS1 to fall appreciably.

Compare the outcome of this analysis to that of the following analysis in which $M_1 = M_2 = 1$. These choices assign equal penalties to both military occupational specialties. The linear programming analysis produced the following results:

Cell (i,j)	P_{ij}	Number of Enlistees Assigned to MOS1	Number of Enlistees Assigned to MOS2
(1,1)	0.00	0	43
(1,2)	1.00	40	0
(1,3)	1.00	25	0
(2,1)	0.00	0	26
(2,2)	0.53	24	21
(2,3)	1	1	0
Total		90	90

$\hat{V}^{new}_{mos1}$=631 and $\hat{V}^{new}_{mos2}$ =.259.

SUMMARY

Three problems associated with the use of military hands-on job performance tests have been addressed in this paper. The first concerned methods for setting standards of minimally acceptable performance on the tests. In addressing that problem, we described standard-setting procedures that have been used in a wide variety of settings in the civilian sector. We then discussed the prospects for using those procedures with the hands-on tests. Finally, we described a set of operational issues that must be addressed, regardless of the standard-setting procedures adopted by the Services. Among the most frequently used standard-setting procedures, those proposed by Angoff (1971) and Nedelsky (1954) appear to hold the greatest promise for use with the performance components and knowledge components, respectively, of the military job performance tests we have reviewed. Examinee-based standard-setting procedures would be most applicable to tests that are not composed of dichotomously scored activities or items.

The second problem we addressed involves procedures for eliciting and combining judgments of the values of enlistees' behaviors on military job performance tests. We examined the potential contributions of psychological decision theory and social behavior theory to solving this problem and concluded that they were largely inapplicable. These theories are more appropriate for eliciting judgments of the values of decision alternatives or

for inferring the attributes of decision alternatives that underlie judges' recommendations. A procedure involving successive lotteries holds promise for defining the values judges attribute to various patterns of enlistees' behavior on military job performance tests.

It appears that all Services have completed extensive job analysis studies and have developed elaborate lists of tasks that compose their military occupational specialties. Additional studies have resulted in the development of taxonomic clusterings of these tasks on such dimensions as frequency, difficulty, and judged importance. The results of these studies can and should be employed in developing methods for combining judged values associated with performance of the tasks that compose a military occupational specialty. A method based on weighted averages of value functions, with weights proportional to the judged importance of tasks, was described in detail.

The third problem addressed in this paper concerns procedures for using enlistees' predicted job performance test scores and judged values associated with those scores in classifying enlistees among military occupational specialties. Several alternatives were considered, including one that considered only the interests and the predicted abilities of individual enlistees (a guidance model) and several that considered only the interests of the Services. Of the latter two, one method presumed that classification decisions were made sequentially, for each individual enlistee. The other method presumed that groups of enlistees were classified concurrently, and that it was desired to effect these classification decisions in a way that maximized the average values of personnel in all military occupational specialties. An explicit solution to this latter problem, in the form of a linear programming algorithm, was described and illustrated.

REFERENCES

Angoff, W.H. 1971 Scales, norms, and equivalent scores. Pp. 508-600 in R. L. Thorndike, ed., *Educational Measurement*. 2nd ed. Washington, D.C.: American Council on Education.

Berk, R.A. 1976 Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education* 45:4-9.

1985 A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests. Paper presented before the annual meeting of the National Council on Measurement in Education, Chicago.

Berk, R.A., ed. 1980 *Criterion-Referenced Measurement: The State of the Art*. Baltimore, Md.: Johns Hopkins University Press.

Burtch, L.D., M.S. Lipscomb, and D.J. Wissman 1982 *Aptitude Requirements Based on Task Difficulty: Methodology for Evaluation* . TR81-34. Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks Air Force Base, Tex.

Committee on the Performance of Military Personnel 1984 *Job Performance Measurement in the Military: Report of a Workshop*. Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, D.C.: National Academy Press.

Cross, L.H., J.C. Impara, R.B. Frary, and R.M. Jaeger 1984 A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement* 21:113-130.

Ebel, R.L. 1972 *Essentials of Educational Measurement*. 2nd ed. Englewood Cliffs, N.J.: Prentice-Hall.

1979 *Essentials of Educational Measurement*. 3rd ed. Englewood Cliffs, N.J.: Prentice-Hall.

Gardiner, P.C., and W. Edwards 1975 Public values: multiattribute-utility measurement for social behavior. Pp. 1-38 in M. F. Kaplan and S. Schwartz, eds., *Human Judgment and Decision Process*. New York: Academic Press.

Glass, G.V 1978 Standards and criteria. *Journal of Educational Measurement* 15:237-261.

Goody, K. 1976 Comprehensive Occupational Data Analysis Programs (CODAP): Use of REXALL to Identify Divergent Raters. TR-76-82, AD-A034 327. Air Force Human Resources Laboratory, Occupation and Manpower Research Division, Lackland Air Force Base, Tex.

Gulliksen, H. 1950 *Theory of Mental Tests*. New York: John Wiley and Sons.

Hambleton, R.K. 1980 Test score validity and standard-setting methods. Pp. 80-123 in R. A. Berk, ed., *Criterion-Referenced Measurement: The State of the Art*. Baltimore, Md.: Johns Hopkins University Press.

Hambleton, R.K., and D.R. Eignor 1980 Competency test development, validation, and standard-setting. Pp. 367-396 in R. M. Jaeger and C. K. Tittle, eds., *Minimum Competency Achievement Testing: Motives*. *Models*. *Measures*, *and Consequences*. Berkeley, Calif.: McCutchan.

Hendrix, W.W., J.H. Ward, M. Pina, and D.D. Haney 1979 *Pre-Enlistment Person-Job Match System*. TR-79-29. Air Force Human Resources Laboratory, Occupation and Manpower Research Division. Brooks Air Force Base, Tex.

Hillier, F.S., and G.J. Lieberman 1974 *Operations Research*. San Francisco: Holden-Day Press.

Jaeger, R.M. 1978 A Proposal for Setting a Standard on the North Carolina High School Competency Test. Paper presented before the annual meeting of the North Carolina Association for Research in Education, Chapel Hill.

1982 An iterative structured judgment process for establishing standards on competency tests: theory and application. *Educational Evaluation and Policy Analysis* 4:461-476.

Jaeger, R.M., and J.C. Busch 1984 *A Validation and Standard-Setting Study of the General Knowledge and Communication Skills Tests of the National Teacher Examinations* . Final report. Greensboro, N.C.: Center for Educational Research and Evaluation, University of North Carolina.

Kaplan, M. 1975 Information integration and social judgment: interaction of judge and informational components. Pp. 139-172 in M. F. Kaplan and S. Schwartz, eds., *Human Judgment and Decision Process*. New York: Academic Press.

Kroeker, L., and J. Folchi 1984a Classification and Assignment within PRIDE (CLASP) System: Development and Evaluation of an Attrition Component. TRØ84-40. Navy Personnel Research and Development Center, San Diego, Calif.

1984b Minority Fill-Rate Component for Marine Corps Recruit Classification: Development and Test. TR 84-46. Navy Personnel Research and Development Center, San Diego, Calif.

Kroeker, L.P., and B.A. Rafacz 1983 Classification and Assignment within PRIDE (CLASP): A Recruit Model. TR 84-9. Navy Personnel Research and Development Center, San Diego, Calif.

Laabs, G.J. 1984 Performance-Based Personnel Classification: An Update. Navy Personnel Research and Development Center, San Diego, Calif.

Livingston, S. A., and M. J. Zieky 1982 *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, N.J.: Educational Testing Service.

1983 *A Comparative Study of Standard-Setting Methods*. Research Report 83-38. Princeton, N.J.: Educational Testing Service.

Lord, F.M., and M.R. Novick 1968 *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.

Meskauskas, J.A. 1976 Evaluation models for criterion-referenced testing: views regarding mastery and standard-setting. *Review of Educational Research* 45:133-158.

Morsch, J.E., J.M. Madden, and R.E. Christal 1961 Job Analysis in the United States Air Force. WADD-TR-61-113, AD-259 389. Personnel Laboratory, Lackland Air Force Base, Tex.

Nedelsky, L. 1954 Absolute grading standards for objective tests. *Educational and Psychological Measurement* 14:3-19.

Poggio, J.P., D.R. Glassnap, and D.S. Eros 1981 An Empirical Investigation of the Angoff, Ebel, and Nedelsky Standard-Setting Methods. Paper presented before the annual meeting of the American Educational Research Association, Los Angeles.

Raimsey-Klee, D.M. 1981 Handbook for the Construction of Task Inventories for Navy Enlisted Ratings. Navy Occupational Development and Analysis Center, Washington, D.C.

Roberts, D.K., and J.W. Ward 1982 General Purpose Person-Job Match System for Air Force Enlisted Accessions. SR-82-2. Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks Air Force Base, Tex.

Schmitz, E.J., and P.B. McWhite 1984 Matching People with Occupations for the Army: The Development of the Enlisted Personnel Allocation System. Personnel Utilization Technical Area Working Paper 84-5. U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Va.

Sherif, M. 1947 Group influences upon the formation of norms and attitudes. In T. M. Newcomb and E. L. Hartley, eds., *Readings in Social Psychology*. 1st ed. New York: Holt.

U.S. Army Research Institute for the Behavioral and Social Sciences 1984 Selecting Job Tasks for Criterion Referenced Tests of MOS Proficiency. Working Paper RS-WP-84-25. U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Va.

Winkler, R.L., and W.L. Hays 1975 *Statistics: Probability*, *Inference*, *and Decision*. 2nd ed. New York: Holt, Rinehart and Winston.

Zedeck, S., and W.F. Cascio 1984 Psychological issues in personnel decisions. *Annual Review of Psychology* 35:461-518.

Zieky, M.L., and S.A. Livingston 1977 *Manual for Setting Standards on the Basic Skills Assessment Tests*. Princeton, N.J.: Educational Testing Service.

Exploring Strategies for Clustering Military Occupations

Paul R. Sackett

CLUSTERING MILITARY OCCUPATIONS

The Joint Services Project on Assessing the Performance of Enlisted Personnel has resulted in the collection of data on a variety of criterion measures for a number of occupational specialties. Intercorrelations among these criteria are being examined, as are relationships between the Armed Services Vocational Aptitude Battery (ASVAB) subtests and composites and performance on these criterion measures. A fundamental issue facing the Services is that of extending the results of these efforts from the limited set of occupational specialties included in the project to the universe of military occupational specialties (MOS).

More specifically, three different types of extension are needed. The first is the issue of ASVAB validity: based on known ASVAB-performance relationships for a small number of MOS, we wish to infer ASVAB-performance relationships for the universe of MOS. The second is the issue of intercorrelations among criteria. For a small number of MOS, intercorrelations among various types of criteria (e.g., hands-on performance tests and training grades) are known; we wish to generalize these relationships to the universe of specialties. The third is the issue of setting predictor cutoffs for various MOS. For MOS for which ASVAB and criterion data are available, it is at least possible (even if not current practice) to set cutoffs to ensure that no more than a specified proportion of applicants will fall below some

specified level of criterion performance. We wish to set justifiable cutoffs for MOS for which high-quality criterion data are not available.

The critical question is what aspects of jobs produce variations in validity coefficients, criterion intercorrelations, and cutoff scores. If this question can be answered, we can then ask two more questions: (1) which MOS can be shown to be sufficiently similar to MOS for which predictor and criterion data are available that we can infer that validity is the same and/or that appropriate cutoffs are the same; and (2) for MOS that are not sufficiently similar to any for which predictor-criterion data are available, can we establish relationships between job characteristics and validity coefficients, criterion intercorrelations, and cutoff scores such that we can make projections to MOS for which predictor-criterion data are not available?

This paper considers approaches to addressing the need to assess job similarity in the context of the questions stated in the above paragraph, rather than as a general review of the job clustering literature. My single greatest concern about both the job analysis and job clustering literatures is the pervasive tendency to ignore the purpose for which job analysis is being done or for which jobs are being compared. When comparing jobs, two major decisions need to be made: (1) what job descriptor to use (e.g., tasks, abilities), and (2) what quantitative clustering procedure to use. The second has received more attention than the first; a detailed review by Harvey (1986) makes it unnecessary to treat this issue in detail here. The first factor has been shown (e.g., Cornelius et al., 1979) to have a large impact on decisions about job similarity. For example, jobs very different at the task level may be quite similar at the ability level. Decisions about the appropriate job descriptor are needed for subsequent efforts to examine the relationships between job characteristics and validities and cutoff scores.

JOB ANALYSIS METHODS: THE CHOICE OF THE JOB DESCRIPTOR

Numerous approaches to analyzing jobs exist. Textbooks in the fields of industrial/organizational psychology and personnel management commonly catalog 6-12 job analysis methods (e.g., functional job analysis, Position Analysis Questionnaire (PAQ), task checklist, job element method, critical incidents, ability requirement scales, threshold traits analysis) (e.g., Cascio, 1982; Schneider and Schmitt, 1986). One way to disentangle the myriad of approaches is to characterize them on a number of dimensions. Dimensions include source of information (e.g., incumbent versus supervisor versus job analyst) and method of collecting information (e.g., observation versus interview versus questionnaire), purpose (e.g., setting selection standards versus setting wages), and job descriptor (e.g., describing tasks versus describing

attributes needed for task performance). Of particular interest here are these last two: the purpose for which the job analysis information is collected and the job descriptor chosen.

Pearlman's (1980) review of the literature on the formation of job families identifies four major categories of job descriptors. The first he labels "job-oriented content," referring to systems that describe work activities in terms of work outcomes or tasks. In other words, the focus is on what work is accomplished. Such systems are job specific. Pearlman gives two examples of task statements: "turns valves to regulate flow of pulp slush from main supply line to pulp machine headbox," and "install cable pressurization systems." I have relabeled this category with the more descriptive title "specific behaviors." Researchers and practitioners describing jobs at this level typically use the label "tasks," and generate a detailed list of tasks statements. Four to five hundred task statements are not uncommon.

Pearlman's second category is labeled "worker-oriented content," referring to systems that describe work activities in terms of behaviors or job demands that are not job specific. Thus these systems are intended as applicable to a wide variety of jobs and commonly involve evaluating jobs using a standard questionnaire. I have relabeled this category "general behaviors." McCormick's Position Analysis Questionnaire (PAQ) typifies this approach. Sample PAQ items include "use quantitative materials" and "estimate speed of moving objects." Thus researchers and practitioners describing jobs in these terms typically use an inventory of 100-200 behavioral statements.

Pearlman's third category is labeled "attribute requirements," referring to systems that describe jobs in terms of the areas of knowledge, skill, or ability needed for successful job performance. Two very different approaches can fall into this category. The first involves the identification of specific areas of knowledge, skill, and ability needed for performance in one specific job in the context of the development of selection tests that will be justified on content validity grounds. This is a very common activity among psychologists developing selection systems in public sector settings. The critical feature is that the applicants for the job in question are expected to already have obtained the training to perform the job; thus the focus is on determining the extent to which applicants possess specific knowledge and skills needed for immediate job performance. In these settings it is not uncommon to develop detailed lists of 100-200 areas of needed knowledge, skill, and ability; these lists are then used to guide test development.

The second is more applicable to the military situation in that it is more applicable to settings in which training takes place after selection. As knowledge and specific skills will be acquired in training, selection is based on abilities shown to be predictive of knowledge and skill acquisition and/or subsequent job performance. Thus, rather than focusing on large numbers of areas of job-specific knowledge and skill, this approach involves describing

jobs in terms of a fixed set of cognitive, perceptual, and psychomotor abilities. I use the label "ability requirements" to refer to this subset of the more general category "attribute requirements." An example of this approach is Fleishman's work on ability requirements (Fleishman and Quaintance, 1984). Based on an extensive program of research, a list of abilities was created, as well as rating scales for evaluating the degree to which each of the abilities is required. The present list identifies 52 abilities; smaller numbers could be used if, for example, motor requirements were not relevant to the purpose for which job information was being collected. A focus solely on cognitive ability requirements would involve 14 abilities. Examples include "number facility" and "fluency of ideas." Thus the ability requirements approach involves describing jobs in terms of a relatively limited number of abilities required for job performance.

Pearlman's fourth category is labeled "overall nature of the job," referring to approaches that characterize jobs very broadly, such as by broad job family (managerial, clerical, sales). An example of this category that may be of particular interest to the Job Performance Measurement Project (JPM Project) is Hunter's (1980) grouping of all 12,000 jobs in the Dictionary of Occupation Titles into one of five categories on a job complexity scale. This complexity scale is based on a recombination of the Data and Things scales used by the U.S. Department of Labor to classify jobs. Hunter shows that validity coefficients for composites of General Aptitude Test Battery (GATB) subtests differ across levels of this complexity variable and are very similar within levels of this variable.

As Pearlman points out, distinctions between these categories are not always clear, and some approaches to job analysis involve multiple categories. However, it is conceptually useful to conceive of these four categories as a continuum from more specific to less specific. A given job can be described in terms of a profile of 400-500 specific behaviors, 100-200 general behaviors, 10-40 abilities, or a single global descriptor, such as job complexity. It should be recognized that this is not merely a continuum of level of specificity; there are clearly qualitative differences in moving from behaviors performed to abilities required. Nonetheless, this discussion should clarify the differences in level of detail involved in the various approaches to describing jobs and should set the stage for a discussion of the relationship between the purpose for which job information is being collected and the type of job descriptor chosen.

These two issues—purpose and job descriptor chosen—are closely intertwined. The question "which job analysis method is most appropriate" can only be answered in the context of a specific purpose. An illustration would be an example from a job analysis of the job "psychologist." An issue of concern was whether different specialties within psychology—clinical, counseling, industrial/organizational, and school—were similar enough that a common licensing

exam was appropriate for these four specialties. The Educational Testing Service (ETS) was commissioned to conduct a comparative job analysis of these four areas (Rosenfeld et al., 1983). An inventory of 59 responsibilities and 111 techniques and knowledge areas was designed and mailed to a carefully selected sample of licensed psychologists. The study found a common core of responsibilities among all four specialties and chided various practice areas for emphasizing the uniqueness of their own group.

I am not denying that there are commonalities among different types of psychologists. However, I will argue that I could have easily designed a survey instrument that would have produced different results. One thing industrial/organizational psychologists have learned from our experience with job analysis is that the more general the data collected, the more likely it is that jobs will appear similar when subjected to statistical analysis; conversely, the more specific the inventory items, the greater the apparent differences among jobs. The art of job analysis lies in determining a level of specificity that meets the purposes of the particular job analysis application. Consider some of the statements making up the ETS inventory. Responsibility 1 leads the inventory reading: "Conduct interviews with client/patient, family members or others to gain an understanding of an individual's perceived problem." This is endorsed by a high proportion of respondents from all specialties, yet it can mean dramatically different things, from interviewing a corporate executive to gain insight into an organization's incentive pay plan to interviewing a 7-year-old suspected victim of child abuse. More examples: "observe the behavior of individuals who are the focus of concern," and "formulate a working hypothesis or diagnosis regarding problems or dysfunctions to be addressed." Again, these can refer to dramatically different activities. More to the point, given that the purpose of the job analysis is to support the creation of one or more licensing exams, these can require different skills, abilities, training and experience. By being more specific and rephrasing Responsibility 1 as multiple tasks ("interview business clients," "interview adult patients," "interview children"), the chances of concluding that the jobs are different increase. By getting even more general ("gather information verbally"), the chances of concluding that the jobs are similar increase. Each of these three levels of specificity present information which is true. However, the question of which level of specificity is appropriate depends on the purpose for which the information is being collected.

In the above example, the three levels of specificity illustrated all focus on worker activities. The job descriptor chosen is in all cases behavioral; they vary on a continuum from general behaviors to specific behaviors. Similarly, one may reach different conclusions about job similarities and differences if different categories of job descriptors are chosen (e.g., focusing on job activities versus focusing on abilities required for job performance).

A multiorganization study of bank teller and customer service jobs illustrates this nicely (Richardson, Bellows, Henry, and Co., 1983). A 66-item behavioral work element questionnaire (e.g., "cashes savings bonds," "verifies signatures," "types entries onto standardized forms") and a 32-item ability requirement questionnaire (e.g., "ability to sort and classify forms," "ability to compute using decimals," "ability to pay attention to detail") were administered. While the vast majority of incumbents held the title "paying and receiving teller," 20 other job titles were found (e.g., new accounts representative, customer service representative, drive-in teller, safe deposit custodian). The issue was whether these 20 jobs were sufficiently similar to the job of paying and receiving teller that a selection test battery developed for the paying and receiving tellers could also be used for the other jobs. A correlation between each job and the paying and receiving teller was computed, first based on the behavioral work element ratings and then based on the ability ratings. In a number of cases, dramatically different findings emerged. The new accounts representative, customer service representative, and safe deposit custodian correlated .21 with the paying and receiving teller when comparing the jobs based on similarity of rated behavioral work elements. These same three jobs correlated .90, .92, and .88 with the paying and receiving teller when comparing the jobs based on similarity of rated ability requirements. Thus the use of different job descriptors leads to different conclusions about job similarity. Conceptually, one could argue that for purposes of developing an ability test battery, the ability requirements data seem better suited. If data on these same jobs were being collected to determine whether a common training program for new hires was feasible, one might argue that the work behavior data seem better suited. Again, the question "which jobs are sufficiently similar that they can be treated the same" cannot be answered without information as to the purpose for which the jobs are being compared.

A study by Cornelius et al. (1979) reinforces this point and takes it one step further. They analyzed seven nominally different first-level supervisory jobs in chemical processing plants. Hierarchical clustering analysis was done to establish job groupings based on three types of data: task similarity, similarity of Position Analysis Questionnaire profiles, and similarity of ability requirements. Each type of data produced a different pattern of job similarities and a different clustering of jobs. Cornelius et al. properly tell us that purpose will dictate which set of data we should rely on. However, even after this decision has been made, problems remain. Cornelius et al.'s task analysis data, for example, indicate that both five-cluster and three-cluster solutions are feasible. Hierarchical cluster analysis, as well as other grouping methods, can only establish *relative* similarity among jobs. In the Cornelius et al. study, if 40 percent of tasks in common is seen as sufficient to label jobs similar, the seven jobs would fall into three clusters.

If 60 percent of tasks in common is seen as sufficient to label jobs similar, the seven jobs would fall into five clusters. The question left unanswered is "given that an appropriate job descriptor has been chosen, how large a difference between jobs on the chosen descriptor is needed to have a significant impact on the criterion of interest?" In a selection setting, how different do jobs have to be before validity coefficients are affected? In a training situation, how different do jobs have to be before separate training programs are required? In a performance appraisal situation, how different do jobs have to be before separate performance ratings forms need to be constructed? Thus job clustering can only be meaningful with reference to an external criterion.

In summary, the above discussion highlights a number of concerns about job grouping. First, different descriptors can produce very different job groupings. Second, different levels of specificity within a given general type of descriptor (e.g., task) can produce very different job groupings. Third, even if a given type of job descriptor and level of specificity are agreed on, the magnitude of job differences that will be needed to classify jobs differently remains a problem. An external criterion is needed.

The implications of the above discussion for the JPM Project are clear. First, there is reason to expect that different job descriptors will produce different job groupings. The choice of job descriptor should not be a function of the availability of job descriptor data using a particular approach, but rather a function of the type of job descriptor data which is most closely linked to the purpose for which jobs are being grouped. Second, it must be realized that the two goals of grouping jobs with similar test validities and grouping jobs with similar levels of ability required to ensure a specified level of performance must be treated independently. Grouping jobs based on validity may produce very different job clusters than grouping jobs based on required ability levels. Conceivably these two purposes could require different job descriptors for optimal clustering. Approaches to identifying the appropriate job descriptor for these purposes are discussed in a subsequent section of this paper.

One additional aspect of the choice of the job descriptor merits some discussion, namely, the nature of the data to be collected about the descriptor chosen. Given that a descriptor has been chosen (e.g., specific behaviors or abilities), it is common to ask job experts to rate the importance of each job component. However, "importance" can be conceptualized in a number of ways, three of which are discussed here. Using abilities as an example, one approach to importance is in terms of time: what proportion of total time on the job is spent using the ability in question. The Position Analysis Questionnaire, for example, uses this type of scale for some items. A second approach is in terms of contribution to variance in job performance: to what extent does the ability in question contribute to differentiating the

more successful employees from the less successful. The job element approach to job analysis for selection system development uses such a scale. A third approach is in terms of level: what degree of a given ability is needed for successful job performance. Fleishman's Ability Requirement Scales exemplify this approach. Conceptually, it is clear that these three can be completely independent. The abilities that are used most frequently may be possessed by virtually all incumbents and thus not contribute to variance in job performance. A given ability may contribute equally to variance in job performance in two jobs, yet the level of ability needed may differ dramatically across the jobs. Thus, even if it were agreed that abilities required is the appropriate job descriptor for a given application, operationalizing ability as importance, frequency of use, contribution to variance in performance, or level required can lead to different conclusions about job similarity. It would seem logical to hypothesize that judgments about contributions to variance in job performance would be most appropriate for determining for which jobs a given test should have similar validity and that judgments about level required would be most appropriate for determining which jobs should have similar test cutoffs.

The distinctions made in the above paragraph are not typically made. In fact, researchers sometimes seem to feel that the choice of the descriptor is all that is important and do not even mention the aspect of the descriptor that is rated. For example, a paper by Cornelius et al. (1984) describes the construction and use of a 26-item ability element battery to group jobs in the petroleum/ petrochemical industry. They used the results of this inventory to assign jobs to one of three occupational groups, but did not tell us whether ability was operationalized as frequency of use, contribution to variance in performance, or level required.

The use of one operationalization of importance where another seems better suited is found in Arvey and Begalla's (1975) examination of the job of homemaker. They administered the PAQ to a sample of homemakers and compared the PAQ profile for this position with each of the large number of profiles in the PAQ data base. These comparisons were made for two human resource management purposes: attempting to associate a wage with the homemaker job and making inferences about job transfer and training decisions. Jobs most similar in PAQ profiles were patrolman, home economist, airport maintenance chief, and kitchen helper; a number of supervisory positions followed closely (electrician foreman, gas plant maintenance foreman, fire captain) in the list of the 20 most similar positions. Arvey and Begalla note that a major theme running through many of the occupations listed was a trouble-shooting emergency handling orientation.

Based on this list of most similar occupations, it is not clear that the goal of identifying jobs amenable to entry by homemakers was met. Arvey and Begalla note this and interpret their findings with appropriate caution. The

predicted salary for the job was $740 per month, in 1969 dollars, which the authors felt was overinflated. They offer distortion of responses based on desire on the part of the respondents to make their positions seem more important as an explanation of the high salary. In light of our discussion of various operationalizations of job element importance, another explanation seems likely: the descriptions provided are accurate (i.e., not intentionally distorted), but the information requested is not well suited to the task at hand. The ratings scales used in the PAQ typically reflect time spent: either a direct rating of frequency or a rating of importance, operationalized vaguely as "consider such factors as amount of time spent, the possible influence on overall job performance if the worker does not properly perform the activity, etc." I would hypothesize that different patterns of similarity would be found if "level required" rather than "time spent" were used to rate items. Conceptually, level required seems better suited to the tasks of identifying jobs amenable to entry by homemakers and setting wage levels. Jobs very similar in the amount of time spent on the PAQ dimension "processing information" may be very different in the level of information processing involved. In short, it is suggested that careful attention be paid to both the selection of the job descriptor and to the operationalization of job element importance.

The following sections of this paper separately address the issues of identifying valid predictors of performance for the universe of MOS and setting minimum standards on these predictors. Multiple potential solutions to the problem are presented.

EXTENDING ASVAB VALIDITY TO THE UNIVERSE OF MOS

Validity Generalization/Meta-Analysis

Validity generalization is a form of meta-analysis. The application of meta-analytic techniques to the examination of predictor-criterion relationship in the selection arena has been labeled validity generalization; the use of the two terms is the result of the parallel development of data cumulation techniques by two groups of researchers—Glass and colleagues (e.g., Glass et al., 1981) and Schmidt and Hunter and colleagues (e.g., Hunter et al., 1982)—who applied different labels to similar techniques. Note that there are five-book length treatments of cumulative techniques (Glass et al., 1981; Hunter et al., 1982; Rosenthal, 1984; Cooper, 1984; Hedges, 1985) and a number of thorough and critical treatments of the topic in archival journals (e.g., Schmidt et al., 1985; Sackett et al., 1985; Bangert-Drowns, 1986).

An introduction to validity generalization is in order. For years psychologists have observed that when a given test is validated in different settings, the resulting validity coefficients vary; in some cases the amount of variation

is substantial. Historically, the explanation offered for this was that situational factors affected validity. Due to these unspecified situational factors (for example, organizational climate, leadership style, and organizational structure) a test valid in one situation might not be valid in another. Thus there is the doctrine of "situation specificity," defined as the belief that due to these factors one could not safely rely on validity studies done elsewhere, but rather, one must do a validity study in each new setting.

To understand validity generalization, it is helpful to distinguish between "true validity" and "observed validity." True validity is the correlation that is obtained if there is an infinitely large sample size that is perfectly representative of the applicant pool of interest and if the criterion measure is a perfectly reliable measure of true job performance. Observed validity is the correlation obtained in our research—typically with smaller Ns than preferred, with samples that may not be perfectly representative of the job applicant population, and with less than perfect criterion measures (e.g., supervisory ratings of performance). Historically, researchers have not differentiated between observed validity and true validity: when observed validity differences were found between studies, it was assumed that the differences were real. Recently, it has been suggested that these differences are *not* real, but simply reflect differences in sample size, criterion reliability, or range restriction. Could it be that true validity does *not* differ across situations? If it weren't for these methodological problems, would validities be the same across studies?

These ideas make for interesting speculation; what was needed were ways of testing them. Validity generalization models are means of testing these ideas: they offer a way of assessing true validity and of assessing how much variability in validity coefficients we can expect due to the methodological problems listed above. The amount of variability in observed validity coefficients is compared with the amount of variability expected due to methodological artifacts: if expected validity equals or nearly equals observed validity, one concludes that differences in validities across studies are not real, but merely the result of the effects of these artifacts.

Procedurally, validity generalization ties together a number of well-known psychometric ideas. One starts with a number of validity studies and a validity coefficient for each. For each study, one obtains an estimate of criterion reliability. Each validity coefficient is corrected using the well-known formula for correction for attenuation in the criterion. Each validity coefficient is also corrected for range restriction—the extent to which the sample used in the study has a narrower range of test scores than would be obtained from job applicants—using well-known formulas for range restriction. The mean and variance of this distribution of corrected validity coefficients is then computed and compared with the variance expected due to sampling error, which is a function of N and the mean validity coefficient. If the

variance expected due to sampling error and the variance in corrected validity coefficients are nearly equal, we conclude that validity is not situation specific, and that the best estimate of true validity is the mean of the corrected validity coefficients.

Validity generalization analyses might appear to be straightforward under the conditions outlined above. However, if criterion reliability values or information about range restriction is not available for each study, assumptions must be made about what criterion reliability was likely to be, about how much range restriction was likely to have occurred, and about the linearity of the predictor/criterion relationship. These assumptions are critical: if the values assumed are incorrect, the estimated value of true validity can be substantially in error. Furthermore, when the range restriction is severe, the extrapolation permitted by these assumptions is tenuous.

A source of confusion in understanding and interpreting validity generalization/meta-analytic research lies in the failure to differentiate between two different statistical tests that can be performed on a set of validity coefficients; these are tests of the situational specificity hypothesis and the generalizability hypothesis. The situational specificity hypothesis is rejected when variance in validity coefficients is essentially zero after correcting for artifacts. Rejecting this hypothesis implies accepting the hypothesis that true validity is virtually constant for the job/test combination under consideration. The generalizability hypothesis is less stringent. It involves the recognition that even if one fails to reject the situational specificity hypothesis and thus acknowledges that validity varies across jobs, it is still possible that even the low end of a distribution of validity coefficients is of a magnitude sufficient to consider the test useful. Thus, if one's interest is not in a point estimate of validity for a given situation but rather in simply the assurance that test validity will be above a level considered minimally acceptable, one can accept the generalization hypothesis if the low end of a confidence interval around mean validity exceeds this level.

The research of Schmidt and Hunter has asserted that cognitive ability tests are valid for all jobs (Hunter and Hunter, 1984). Some have interpreted this as implying that tests are *equally* valid for all jobs. This misinterpretation is based on confusing the situational specificity hypothesis and the generalizability hypothesis. Schmidt and Hunter's statements involve accepting the generalizability hypothesis, (i.e., that the validity of cognitive tests is positive and nonzero for all jobs).

While validity generalization research with cognitive ability tests shows quite strongly that there is little to no variation in true validity for individual job/test combinations, it is very clear that the validity of cognitive ability tests does vary across jobs. One of the clearest illustrations of this is found in a study by Schmidt et al. (1981). For a sample of 35 Army jobs, validity coefficients for the 10 subtests of the Army Classification Battery

were available for two independent samples of about 300 individuals per job. For each subtest, the 35 validity coefficients from the first sample were correlated with the 35 validity coefficients from the second sample. With the exception of one subtest (radiocode aptitude), correlations between samples were substantial, ranging from .68 to .86. The pattern of validity coefficients was stable across samples: jobs with higher validities in one sample had high validities in the other sample, and vice versa. If true validity did not vary across jobs, variation from sample to sample would be a function of sampling error, and the correlation across samples would be essentially zero. Thus, jobs do moderate validity, if by "moderate" we mean "influence the size of a validity coefficient." However, in this study, Schmidt et al. define "moderate" as "produce a near zero validity." Using this definition. they conclude that jobs do not moderate validity, as the low end of a confidence interval (i.e., two standard deviations below the mean validity) is greater than zero. This formulation accepts the generalization hypothesis, and by definition rejects the situational specificity hypothesis. A less extreme definition of "moderate" would lead one to support both hypotheses.

This somewhat lengthy introduction to validity generalization sets the stage for considering the application of validity generalization to the JPM Project. Assume the availability of validity coefficients for ASVAB subtests and Service-wide composites for each of the 27 MOS included in the project, using hands-on performance tests as the criterion. For each subtest and composite, observed and expected variance can be computed and compared, and residual variance can be used to put a confidence interval around the mean of the validity coefficients. If this lower bound is positive and nonzero, it has thus been shown that the test in question is predictive of job performance for the MOS in question. If one feels confident that the sampled MOS are representative of the universe of MOS, this conclusion is generalized to the universe.

A number of comments on this approach are needed. First, offering this as at least a partial solution to the question of demonstrating ASVAB validity for predicting on-the-job performance is contingent on producing the expected results, namely, that the lower bound for validity will prove to be positive nonzero. The body of research leading to the expectation that this result will be found is substantial (see Hunter, 1980; Hunter and Hunter, 1984). The one potentially important difference between the present set of validity studies and the cumulated literature on the validity of cognitive ability tests is the criterion used. Most validity generalization work to date has categorized studies as using training criteria (typically end-of-course knowledge test scores) or performance criteria (typically supervisory ratings) (Hunter and Hunter, 1984). Is there reason to expect a different pattern of results using hands-on job performance criteria? Recent work by Hunter (1986) suggests not. Hunter found 12 studies where three different

types of criteria were collected: hands-on work samples, paper and pencil job knowledge tests, and supervisory ratings. Breaking these studies down into military and civilian subsamples, he found that general cognitive ability, after correcting for restriction of range and criterion unreliability, correlated .80 with knowledge, .75 with work samples, and .47 with ratings in the civilian subsample, and .63 with knowledge, .53 with work samples, and .24 with ratings in the military subsample. Knowledge and work sample criteria correlated .80 and .70 in the civilian and military subsamples, respectively, suggesting that a high degree of similarity between validity findings using knowledge criteria and work sample criteria is likely. However, recently completed and as yet unpublished research undertaken as part of the JPM Project indicates lower levels of validity using hands-on performance measures.

Second, this approach presumes that it is sufficient to demonstrate positive nonzero validity; point estimates of true validity are not necessary. As discussed above, it is clear that the true validity of cognitive ability tests does vary across jobs; if one wishes to estimate true validity for MOS not included in the Job Performance Measurement Project, a system of relating variance in job descriptors to variance in validity coefficients is needed. Approaches to such a system will be discussed in the section below on synthetic validity.

Third, the validity generalization approach discussed here is directly relevant only to the issue of establishing test validity and not to the issue of setting selection standards for various jobs. Both of these issues could be dealt with simultaneously with a validity generalization model dealing in regression slopes and intercepts rather than correlation coefficients; in fact, such a model has been developed by Raju (1986). However, this type of model is only applicable in situations in which a common predictor and criterion metric are used in all studies. Thus, such a model might be applied with a single organization if, for example, the job performance of sales clerks was measured using the same procedure in 20 retail stores and regression equations were computed for each store. In the JPM Project, as in most validity generalization applications, the criterion metric varies across studies. Standardizing the data within each organization prior to cumulation is not a solution: the resulting standardized beta weight is the correlation. Note, though, that at one level the issue of justifying cutoffs can be addressed. If a test is valid and the relationship between test and criterion is linear, it can be argued that any cutoff is justifiable in the sense that there is no single point above which individuals will perform successfully and below which individuals will not perform successfully. Extensive research by Hawk (1970) shows that test-criterion relationships for cognitive ability tests do not depart from linearity at a rate greater than would be expected by chance. Any chosen cutoff is justifiable in the sense that individuals above the

cutoff have a higher probability of success than individuals below the cutoff. Thus it could be argued that the issue of "validating" a cutoff score is not intrinsically meaningful, and supply and demand and judgments by policy makers about the relative importance of various MOS can be the basis for establishing cutoffs.

Fourth, it should be noted that some skepticism about validity generalization remains. Some of this is naive. For example, "just because a test predicts performance in these twenty settings is no guarantee that it will predict performance in the twenty-first setting; therefore a local study is needed." By this logic, the local study is also useless: just because the test predicts performance in the validation sample is no guarantee that it will do so with new applicant samples. Some is more sophisticated, such as concerns about the correction of validities for range restriction based on assumed rather than empirically determined measures of the degree of range restriction, or concerns about the statistical power of validity generalization procedures when applied to small numbers of validity coefficients (cf., Sackett et al., 1985). However, one indication of the degree of acceptance of validity generalization can be found in the 1987 *Principles for the Validation and Use of Personnel Selection Procedures* published by the Society for Industrial and Organizational Psychology, Division 14 of the American Psychological Association: "Current research has shown that the differential effects of numerous variables are not so great as heretofore assumed; much of the difference in observed outcomes of validation research can be attributed to statistical artifacts. . . . it now seems well established from both validity generalization studies and cooperative validation efforts that validities are more generalizable than has usually been believed" (p. 26).

The careful sampling of the full spectrum of MOS provides a basis for more confidence than one would usually have in conducting meta-analyses on 27 effect size measures. Recent work on the statistical power of meta-analysis to detect existing moderator variables (Sackett et al., 1986) indicates that a meta-analysis of 27 effect-size measures with average sample sizes of about 150 will be quite powerful.

Linking Hands-On Performance Measures and Training Criteria

A second application of meta-analytic techniques may be appropriate for this project. As discussed earlier, Hunter (1986) examined the relationship between hands-on performance measures and paper and pencil job knowledge tests and found an average correlation of .80 in civilian samples and .70 in military samples. Hands-on performance measures will be available for all 27 MOS in the JPM Project; if training performance is retrievable for the subjects in these samples, this finding can be replicated. Correlations between hands-on measures and training grades can be computed for each

sample and used as input for a meta-analysis. Should the lower bound of this distribution of correlations be reasonably high, we can have some confidence that correlations between ASVAB and training can be generalized to on-the-job performance. Hunter et al. (1985) summarize the substantial body of data relating ASVAB scores to training grades; confirming a strong relationship between training criteria and hands-on criteria could serve as a partial response to a critic concerned about the validity generalization analyses discussed earlier on grounds that only a limited set of MOS were actually included in the analyses. A linkage to a much larger body of literature would thus be made. Another possibility is to correlate the validity coefficients of ASVAB and training with the validity coefficients of ASVAB and hands-on measures. Such a correlation should be interpreted carefully however: unless there is meaningful nonartifactual variance in both distributions of validity coefficients, a relationship between the two sets of validity coefficients can not be obtained.

Synthetic Validity

The concept of synthetic validity is not new. The basic notion is that if various job components can be identified and the validity of predictors of performance on jobs involving each component can be established, one can identify valid predictors of performance in a new job if one knows which job components constitute that new job. A wide variety of techniques have been proposed and/or examined under the rubric of synthetic validity. Trattner (1982) identified four different synthetic validity models: Lawshe's synthetic validity (Lawshe, 1952), Guion's synthetic validity (Guion, 1965), McCormick's job component validity (McCormick et al., 1972), and Primoff's J-coefficient (Primoff, 1975). Mossholder and Arvey's (1984) review of synthetic validity approaches noted that synthetic validity has been talked about substantially more often than it has been applied. Mossholder and Arvey singled out McCormick's job component model and Primoff's J-coefficient as two approaches that have been the focus of serious research efforts. This paper examines the applicability of these two approaches to the present problem of establishing validity for new MOS.

McCormick's Job Component Model

In this approach, the unit of analysis is the job. For a number of jobs, validity coefficients for a given predictor/criterion combination are obtained. Information about each job is obtained through a structured job analysis questionnaire; job dimension scores are derived from this questionnaire and then used as predictors of the validity coefficients for each job. Thus, this is a logical follow-up to validity generalization analysis: for predictor/criterion

combinations for which validity coefficients are found to exhibit more variance than would be expected as a result of artifact, job dimensions identified through structured job analysis are examined as possible moderators of the validity coefficients.

The job components used by McCormick are derived from the Position Analysis Questionnaire (PAQ), a 187-item structured worker-oriented job analysis instrument (McCormick et al., 1972). Factor analysis of the PAQ has produced 32 dimensions, or components; these can be further reduced to 13 overall dimensions. Mecham et al. (1977) identified 163 jobs for which both PAQ ratings and validity coefficients for each of the nine General Aptitude Test Battery (GATB) subtests were available. For each GATB subtest they regressed test validity on the PAQ dimensions. Results were disappointing: shrunken multiple correlations were near zero for four tests, in the teens for two tests, in the .20s for two tests (intelligence and spatial aptitude), and .39 for manual dexterity. They conducted similar analyses using mean test score rather than validity coefficients as the criterion with much more success; these analyses will be discussed in a subsequent section dealing with setting cutoff scores.

A similar approach was taken by Gutenberg et al. (1983). In contrast to the raw empiricism of Mecham et al., Gutenberg et al. hypothesized that specific PAQ dimensions would moderate the validity of specific GATB subtests. They found that two PAQ dimensions, decision making and information processing, correlated significantly with cognitive GATB subtests.

The correlations between job dimensions and validity coefficients obtained in the two GATB studies have not been as large as one might hope for. However, it should be noted that the Gutenberg et al. study corrected validity coefficients for range restriction and produced larger correlations than Mecham et al., suggesting that methodological artifacts may be constraining the relationship between job dimensions and validity coefficients. Note that sampling error has accounted for most artifactual variance in meta-analytic studies and that Gutenberg et al. report that sample sizes for the 111 jobs used in their study ranged from 31 to 537. A reanalysis of their data to determine the impact of sample size would be informative. A median split could be made based on sample size and the analyses repeated separately for the group of studies with relatively large sample sizes and the group with relatively small sample sizes. Assuming that sample size is not systematically associated with some job dimension, we would expect a substantially larger relationship between job dimensions and validity coefficients in the large sample size group; this should be a better estimate of the degree to which job dimensions moderate GATB validities.

If validity coefficients for a given predictor/criterion combination can be predicted from PAQ job dimensions, validity for new MOS could be established

by obtaining PAQ ratings for the new MOS and applying the appropriate prediction formula. An immediate drawback in this approach is the availability of validity data for only 27 MOS. Obviously the Mecham et al. approach of using all PAQ dimensions in a regression equation is not feasible: 45 predictors and 27 cases precludes such an approach. More viable is the Gutenberg et al. approach of identifying a small number of job dimensions on a priori grounds for each predictor/criterion combination under consideration. A panel of psychologists could be asked to reach consensus on the five dimensions most likely to moderate validity coefficients for each predictor/ criterion combination. Regression equations using these five predictors could be computed using, say, 20 of the 27 MOS included in the JPM Project; these equations could then be applied to each of the seven holdout MOS as a test of the effectiveness of the procedure for estimating validity for new MOS. Implicit in the above discussion is the need to obtain PAQ profiles on each of the 27 MOS.

While the above discussion focused on using PAQ dimensions as the job descriptor, the approach outlined above could be undertaken using any standardized job descriptor. One possible explanation for McCormick et al.'s lack of success in predicting validity coefficients using PAQ dimensions is that PAQ dimensions do not constitute the most appropriate job descriptor for this purpose. Consider the array of job descriptors discussed in an earlier section of this paper: specific behaviors, general behaviors, ability requirements, and global descriptors. Issues related to the use of each for this purpose will be reviewed.

One issue is practicality. This synthetic validity model requires a standardized job descriptor system applicable to all MOS. Thus, a system describing each MOS in terms of job-specific behaviors cannot be used in this approach. The availability of data on only 27 MOS also imposes practical constraints. It was proposed above that if the PAQ were used as the job descriptor, expert judgment would be used to identify a subset of PAQ dimensions for examination as possible moderators of validity. This is clearly a makeshift approach, and the possibility that the optimal dimensions will not be selected is very real. This problem is minimized or eliminated if the job descriptor system used involves a small number of dimensions. Selecting 5 of 10-15 abilities used in an ability requirement approach seems less likely to exclude critical dimensions than selecting 5 of 45 PAQ dimensions. Using a global descriptor, such as Hunter's job complexity scale, eliminates the problem entirely.

Another issue is the conceptual appropriateness of each type of job descriptor for this purpose. This discussion can be avoided and replaced by brute empiricism: for each of the 27 MOS included in the project, job analytic work could be done to produce job profiles in terms of general behavioral dimensions, ability dimensions, and global descriptors. The extent

to which each of these factors moderates validity could be examined. However, there is some basis for predicting the outcome of such an effort. First, the validity generalization literature discussed earlier has led to the recognition that within a class of jobs, such as clerical work, differences in specific behaviors performed do not have a substantial influence on validity. Commonality of underlying abilities required leads to similar validity despite lack of overlap in specific behaviors performed. This leads to the hypothesis that more general approaches, namely, ability requirements or global descriptors, are better candidates. Second, successful attempts at examining moderators of validity across diverse jobs have used general rather than molecular job descriptors. Hunter (1980) found that regression weights for using a general cognitive ability composite to predict performance increased from .07 to .40, moving from the lowest to the highest levels of his job complexity scale in a sample of 515 GATB validity studies; similarly, the regression weights of psychomotor ability decreased from .46 to .07, moving from the lowest to the highest levels of the complexity scale. The PAQ dimensions used successfully by Gutenberg et al. (decision making and information processing) are among the most "abilitylike" of the PAQ dimensions.

Third, as Pearlman (1980) notes, the more molecular approaches lack the isomorphism with the individual differences variables being considered as predictors of performance that is found with the molar approaches. Isomorphism between job descriptor constructs and predictor constructs is conceptually elegant, making for a readily explainable and interpretable system. Thus, while isomorphism is by no means a requirement for a successful approach to examining moderators of validity, it is certainly a virtue if such an approach proves viable. Pearlman suggests the use of ability requirements as the descriptor to be used for job grouping for validity purposes.

Therefore, I would suggest that ability requirements and global job complexity be considered as moderators of validity. Fleishman's ability requirement scales (Fleishman and Quaintance 1984) seem particularly worthy of consideration due to the extensive research leading to the development of the scales and the care taken in the definition of each ability. A separate rating scale is provided for each ability, containing a general definition of the ability, definitions of the high and low ends of the scale, a description of how the ability differs from other abilities, and illustrative tasks for various levels of the ability. For example, low, medium, and high levels of the ability "verbal comprehension" are illustrated by "understand a comic book," "understand a newspaper article in the society section reporting on a recent party," and "understand in entirety a mortgage contract for a new home."

Recall our earlier discussion of possible operationalizations of the importance of a given ability as time spent using the ability, contribution of the ability to variance in performance, and level of the ability required. The

Fleishman scales clearly fall into the third category. Conceptually, this third category—level required—seems better suited as a moderator of predictor cutoffs than of validity. The second—contribution to variance in performance—seems better suited to the task at hand. Thus, a separate importance rating, explicitly defining importance as contribution to variance, might be obtained along with the level required rating.

Therefore, it is suggested that ratings of each of the 27 project MOS be obtained using the Fleishman scales with the modification discussed above. Ratings should be made by a number of independent raters to achieve adequate reliability. Existing task analyses of these MOS should aid the rating process. If rated ability requirements are found to moderate validity, predictions of validity for new MOS can then be made.

J-coefficient

A wide variety of algebraically equivalent versions of the J-coefficient are available (see Hamilton, 1981; Primoff, 1955; Urry, 1978). Trattner (1982) describes the J-coefficient as the correlation of a weighted sum of standardized work behavior scores with a test score. Exactly what constitutes these standardized work behaviors, or job elements, varies across J-coefficient applications. In other words, the J-coefficient is a means of estimating the correlation between a test and a composite criterion. The computation of a J-coefficient for a given predictor requires (1) the correlation between the predictor and each criterion dimension, (2) intercorrelations among criterion dimensions, and (3) importance weights for each criterion dimensions.

It is critical to note that the importance of various criterion constructs is a policy issues as well as a scientific one, and take issue with the notion that there is such a thing as "true" overall performance. Consider, for example, two potential dimensions of military job performance: current job knowledge and performance under adverse conditions. It does not seem unreasonable that a policy directive to emphasize combat readiness would increase the importance attached to the second relative to the first. Presuming a lack of perfect correspondence between individuals' standing on the two criterion constructs, the rank order of a group of individuals on a composite criterion would change; which order is "right" reflects policy priorities. Thus the scientific contribution is to identify predictors of each criterion construct; for any given set of weighted criteria we can then estimate the validity of a selection system.

The relevance of the J-coefficient to this project lies mainly in the contribution the approach can make to the issue of determining the correlation between each predictor and each criterion construct. The J-coefficient formulas, of course, accept any validity estimate; users of the approach typically

rely on judgments of the relevance of test items or entire tests for each criterion construct. As judgmental approaches to validity estimation will be discussed separately, no further attention is needed for the J-coefficient itself.

Judgmental Estimates of Validity

Recent research has reported considerable success in obtaining validity estimates by pooling the direct judgments of test validity across a number of judges. Schmidt et al. (1983) and Hirsh et al. (1986) asked psychologists to provide direct estimates of the validity of six subtests of the Naval Basic Test Battery, using training performance as the criterion, for a set of nine jobs. The jobs were selected because of the availability of criterion-related validity studies with sample sizes greater than 2,000, thus providing a standard against which the judgments could be compared that was virtually free of sampling error. Schmidt et al. used experienced psychologists as judges and found that the pooled judgment of ten psychologists deviated on average from the true value by the same amount as would be expected in a criterion-related validity study with a sample size of 673. In other words, this pooled judgment provided a far better estimate of validity than all but the largest scale validity studies. In contrast, Hirsh et al. used the same job/ test combinations with a sample of new Ph.D.s and found that the judgment of a single experienced Ph.D. was as accurate as the pooled judgment of ten new Ph.D.s. The pooled judgment of ten new Ph.D.s proved as accurate as a validity study with a sample size of 104.

The differences found between experienced and inexperienced judges are of great interest. Schmidt et al. (1983) attribute the success of experienced judges to their experience conducting validation research and in accumulating information about validity research done by others. This line of reasoning suggests that even experienced judges will not be successful in estimating validity for predictor/ criterion combinations for which little prior information is available. An alternative explanation for the success of experienced judges is that it is simply due to broader experience with the world of work. They have spent more time in the workplace and have better insights into job requirements. Thus, even for predictor/criterion combinations for which no validity evidence exists at present, they may be able to make accurate judgments. Note that in the J-coefficient literature there is evidence that job incumbent judgments of test-criterion relationships are predictive of empirical validity results, suggesting that work experience, rather than test validation experience, may be the critical factor. Thus there is some basis for positing both that experienced psychologists will be able to estimate validity for a wide variety of predictor-criterion combinations and that experienced nonpsychologists, such as job incumbents,

may also be able to do so. Panels of psychologists and experienced incumbents could be assembled and asked to make validity judgments. As nonpsychologists are not likely to be comfortable with estimating validity coefficients, ratings of the importance of the predictor construct for differentiating between high and low performers on the criterion construct could be obtained and correspondence between these ratings and empirical validity coefficients determined empirically.

Schmidt et al.'s (1983) speculation that the success of experienced psychologists is a function of their memory of validation results for other jobs could be examined. Both the psychologist and incumbent samples could first be asked to estimate validity for five MOS included in the Job Performance Measurement Project in the absence of any information about project results and then asked to estimate validity for five additional MOS. For these additional MOS the judges will be presented with Job Performance Measurement Project validity results for the other 22 MOS to serve as anchors for their judgments. Thus, the impact of information about predictor-criterion relationships for other specialties on the accuracy of validity judgments could be examined.

Paired Comparison Judgments of Validity

An alternative approach to estimating validity judgmentally is the use of paired comparison judgments. Rather than estimating validity directly, judges could be presented with pairs of occupational specialties and asked, for each predictor-criterion combination, which specialty in the pair has the higher validity coefficient. Paired comparison judgments could be obtained from psychologists for 20 of the 27 MOS in the Job Performance Measurement Project data base. These judgments could be pooled across raters and scaled, and the scaling solution then compared with obtained validity coefficients from the project. If a substantial degree of correspondence was found between the scaling solution and obtained validity coefficients, then validity estimates for new MOS could be produced by obtaining paired comparison judgments comparing the new MOS with those for which validity is known and thus mapping the new MOS into the scaling solution. The seven holdout MOS could be used to demonstrate the viability of this approach. This approach is dependent on the assumption that the JPM Project data base includes the full range of MOS, such that the scale points represent the full range of validity coefficients likely to be obtained. Note that this approach demands that judges be very knowledgeable of all MOS involved in the judgments. However, a complete set of judgments is not needed from each judge; each judge can rate partially overlapping sets of MOS. Finally, note that with modification of existing software, this judgment task can be administered by computer.

Determination of Minimum Standards

The determination of minimum cutoff scores has been and remains a problem for which no simple or agreed upon solution exists. Many approaches to setting cutoffs have been identified (e.g., Buck, 1977; Campion and Pursell, 1980; Drauden, 1977; Nedelsky, 1954; Guion, 1965.) One thing all have in common is a subjective component: setting a cutting score requires a value judgment (Cronbach, 1949).

Much of the discussion of cutoff scores is in the context of either achievement testing in an educational setting or the use of content valid work sample or knowledge tests in public sector employment settings. In both of these settings one is typically setting a predictor cutoff in the absence of criterion information. Judgments about expected test item performance of minimally satisfactory performers are typically combined to identify minimum test cutoffs. Without criterion data, a standard is lacking for which these techniques for setting cutoffs can be evaluated. With criterion data and a large sample size, a different type of approach is possible. Based on expert judgment, the minimum acceptable level of criterion performance is identified, and the regression equation relating predictor and criterion is used to identify the predictor score corresponding to this minimum level of acceptable performance. Given the probabilistic nature of predictor-criterion relationships, some individuals scoring above this cutoff will fail and some individuals scoring below this cutoff will succeed. The relative value assigned by the organization to each of these types of prediction error will influence the choice of actual cutoff.

This approach could be applied to all predictor-criterion combinations for the 27 MOS included in the project. Panels of officers directly supervising individuals in each of the 27 MOS could be convened to reach consensus on the minimum acceptable level of performance on each performance construct. Two panels of five for each MOS would provide a group size conducive to consensus decision making and allow a comparison of the judgments of two independent panels. Thus for each predictor-criterion combination for each MOS the predictor score corresponding to minimum acceptable criterion performance could be identified. The availability of these predictor scores would provide a standard for evaluation against which techniques for setting cutoff scores can be assessed even in situations in which empirical predictor and criterion data are not available (e.g., new MOS).

Three techniques for identifying predictor cutoffs are examined. These directly parallel techniques proposed for estimating validity for new MOS. Each is discussed in turn.

Earlier we discussed the use of a synthetic validity approach to examining variance in validity coefficients across MOS. Examples of this approach using the PAQ were examined, and recommendations were made that ability

requirements, rather than general or specific job behaviors, be used as the standardized job descriptor. In applying this approach to setting cutoff scores, previous research has reported a high degree of success (median correlation of .73) in using PAQ dimension scores as predictorss of mean test scores obtained by job incumbents (Mecham et al., 1977.) The Mecham et al. work is based on what they call the "gravitational hypothesis," namely, that people gravitate to jobs commensurate with their abilities. Mecham et al. advocate cutoff scores based on predicted mean test score (e.g., a cutoff of 1 or 2 standard deviations below the predicted mean). This approach is conceptually meaningful only when individuals are free to gravitate to particular jobs. If a test is used to assign individuals to jobs, mean test score merely reflects the organization's a priori assumptions about the job, rather than revealing anything about needed ability levels. Thus, in the military context, predicting mean test score is not very informative. However, the general strategy of using job information (e.g., PAQ dimensions) to predict needed predictor construct scores can be applied by substituting the regression-based predictor score corresponding to the needed minimum level of criterion performance for the mean predictor score.

Again, given 45 PAQ dimensions and 27 MOS, judgments of the PAQ dimensions most likely to be predictive of the needed predictor construct scores could be obtained for each predictor construct to achieve a reasonable ratio of predictors to cases. For each predictor construct, regression equations using 5 PAQ dimensions to predict variance in needed predictor construct scores could be computed for 20 MOS; the resulting equations will be applied to the 7 holdout MOS.

As was the case in considering moderators of validity coefficients, alternatives to the PAQ as the job descriptor of choice should be considered. Without repeating the earlier discussion, the Fleishman ability scales seem particularly well-suited to this task. The explicit measurement of level of ability required links directly to the task of setting predictor cutoffs.

The second approach involves direct estimates of minimum predictor cutoffs. While many approaches to setting cutoffs are based on judgments about predictors (Buck, 1977), such approaches typically involve judgments at the test item level (e.g., judged likelihood that each response option will be chosen by minimally qualified applicants). Such approaches are conceptually more meaningful when dealing with achievement tests, such as those used in an educational setting, than with ability, interest, or biodata measures. Thus, rather than aggregating item-level judgments, direct judgments of minimum predictor cutoffs could be examined. As in the case of direct estimates of validity, panels of psychologists and incumbents could be convened to estimate needed cutoff scores for 5 MOS in the absence of information about cutoff scores for other MOS and then make estimates for 5 additional MOS with access to the regression-based predictor cutoffs for the other 22 MOS.

Finally, a paired comparison process similar to that proposed for validity estimation could be examined. Psychologists could be asked to judge which of a pair of MOS requires a higher predictor score for a given predictor-criterion combination. Judgments could be obtained for all pairs for 20 project MOS; these judgments could be scaled and compared with the regression-based predictor cutoffs. Each of the 7 holdout MOS could then be compared with the MOS for which cutoffs are known and the results mapped into the scaling solution to produce cutoff estimates for the holdout MOS.

CONCLUSION

In retrospect, this paper may be mistitled. The focus has not been on clustering per se, but rather on exploring possible approaches to extending validity findings and empirically based predictor cutoffs beyond the 27 MOS included in the Job Performance Measurement Project. No single best approach has been identified; rather, a number of possibilities have been examined.

A critical question is whether point estimates of validity are needed for various MOS, or whether all that is needed is confidence that the predictors in question have meaningful levels of validity for various MOS. If the second will suffice, the dual strategy of conducting a meta-analysis of the validity studies correlating ASVAB subtests and composites with hands-on performance measures and conducting a meta-analysis of hands-on performance-training performance correlations should provide a clear picture of ASVAB validity for the universe of MOS. The analysis of correlations between ASVAB and hands-on measures is expected, at least by this author, to produce a similar pattern of findings to meta-analyses of cognitive ability tests using training or rating criteria; the expected strong relationship between hands-on measures and training criteria provides a link to the larger body of validity studies using training criteria.

If point estimates of validity are needed, a number of possibilities have been proposed: synthetic validity, direct estimation of validity, and paired comparison judgments of job similarity. Each could be attempted and the relative validity, cost, and ease of use of each could be examined. Considerable attention was paid to the issue of the choice of job descriptor, as the synthetic validity approach involves regressing validity coefficients on standardized job descriptors. Conceptual arguments as well as empirical data were reviewed dealing with the choice of specific behavior, general behavior, ability requirements, and global job information as the job descriptor. While the choice can be viewed as an empirical question to be answered by analyzing the 27 MOS involved in the project using multiple job analytic systems, a strong argument was made for using general, rather than molecular job descriptors, with particular attention paid to ability requirements as

the descriptor of choice. Each of these three approaches to generalizing point estimates of validity was seen as applicable with minor modification to the issue of establishing predictor cutoffs.

As indicated in the opening section of this paper, attention has not been paid to quantitative procedures for grouping jobs. The concern with both the descriptive and inferential grouping methods was that groupings were made on the basis of relative similarity of jobs. What was lacking was an external criterion for determining whether jobs were sufficiently similar that they could be treated the same for the purpose at hand. Data showing that job A was more similar to job B than to job C is not useful without a basis for knowing whether or not the magnitude of differences between the jobs is enough to require that the jobs be treated differently. The synthetic validity approaches discussed in this paper offer the needed criterion. The magnitude of differences on an ability requirement scale needed to produce a change in cutoff score of a given magnitude can be determined, and then used to guide clustering decisions. Hierarchical clustering procedures produce a full range of possible clustering solutions, from each job as an independent cluster to all jobs grouped in one large cluster. At each interim stage, the size of within-cluster differences can be determined; with information as to the magnitude of differences needed to affect the personnel decision in question, one has a basis for informed decisions as to the appropriate number of clusters to retain and as to which jobs can be treated the same for the purpose at hand.

REFERENCES

Arvey, R.D., and M.E. Begalla 1975 Analyzing the homemaker job using the Position Analysis Questionnaire. *Journal of Applied Psychology* 60:513-517.

Bangert-Drowns, R.L. 1986 Review of developments in meta-analytic method. *Psychological Bulletin* 99:388-399.

Buck, L.S. 1977 Guide to the setting of appropriate cutting scores for written tests: a summary of the concerns and procedures. Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center.

Campion, M.A., and E.D. Pursell 1980 Adverse impact, expected job performance, and the determination of cut scores. Paper presented at the meeting of the American Psychological Association, Montreal, August.

Cascio, W. 1982 *Applied Psychology in Personnel Management*. Reston, Va.: Reston Publishing.

Cooper, H.M. 1984 *The Integrative Research Review*. Beverly Hills, Calif.: Sage Publications.

Cornelius, E.T., T.J. Carron, and M.N. Collins 1979 Job analysis models and job classification. *Personnel Psychology* 32:693-708.

Cornelius, E.T., F.L. Schmidt, and T.J. Carron 1984 Job classification approaches and the implementation of validity generalization results. *Personnel Psychology* 37:247-260.

Cronbach, L.J. 1949 *Essentials of Psychological Testing*, 3rd. ed. New York: Harper and Row.

Drauden, G. 1977 Setting Passing Points. Minneapolis Civil Service Commission, Minneapolis, Minn.

Fleishman, E.A., and M.K. Quaintance 1984 *Taxonomies of Human Performance*. Orlando, Fla.: Academic Press.

Glass, G.V., B. McGaw, and M.L. Smith 1981 *Meta-analysis in Social Research*. Beverly Hills, Calif.: Sage Publications.

Guion, R.L. 1965 *Personnel Testing*. New York: McGraw-Hill.

Gutenberg, R.L., R.D. Arvey, H.G. Osburn, and P.R. Jeanneret 1983 Moderating effects of decision-making/information-processing job dimensions on test validities. *Journal of Applied Psychology* 68:602-608.

Hamilton, J.W. 1981 Options for small sample sizes in validation: a case for the J-coefficient. *Personnel Psychology* 34:805-816.

Harvey, R.J 1986 Quantitative approaches to job classification: a review and critique. *Personnel Psychology* 39:267-289.

Hawk, J. 1970 Linearity of criterion-GATB aptitude relationships. *Measurement and Evaluation in Guidance* 2:249-251.

Hedges, L.V. 1985 *Statistical Methods for Meta-analysis*. New York: Academic Press.

Hirsh, H.R., F.L. Schmidt, and J.E. Hunter 1986 Estimation of employment validities by less experienced judges. *Personnel Psychology* 39:337-344.

Hunter, J.E. 1980 Test validation for 12,000 jobs: an application of synthetic validity and validity generalization to the GATB. Washington, D.C.: U.S. Employment Service, U.S. Department of Labor.

1986 Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior* 29:340-362.

Hunter, J.E., J.J. Crosson, and D.H. Friedman 1985 The Validity of the ASVAB for Civilian and Military Job Performance. Rockville, Md.: Research Applications, Inc.

Hunter, J.E., and R.F. Hunter 1984 Validity and utility of alternative predictors of job performance. *Psychological Bulletin* 96:72-98.

Hunter, J.E., F.L. Schmidt, and G.B. Jackson 1982 *Meta-analysis: Cumulating Research Findings Across Studies*. Beverly Hills, Calif.: Sage Publications.

Lawshe, C.H. 1952 What can industrial psychology do for small business? *Personnel Psychology* 5:31-34.

McCormick, E.J., P.R. Jeanneret, and R.C. Mecham 1972 A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire. *Journal of Applied Psychology Monograph* 56:347-368.

Mecham, R.C., E.J. McCormick, and P.R. Jeanneret 1977 *Position Analysis Questionnaire Technical Manual*, *System II*. PAQ Services, Inc. (Available from University Book Store, 360 W. State St., West Lafayette, IN 47906.)

Mossholder, K.W., and Arvey, R.D. 1984 Synthetic validity: a conceptual and comparative review. Journal of Applied Psychology 69:322-333.

Nedelsky, L. 1954 Absolute grading standards for objective tests. *Educational and Psychological Measurement* 14:3-19.

Pearlman, K. 1980 Job families: a review and discussion of their implications for personnel selection. *Psychological Bulletin* 87:1-28.

Pearlman, K., F.L. Schmidt, and J.E. Hunter 1980 Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology* 65:373-406.

Primoff, E.S. 1955 Test Selection by Job Analysis (Technical test series, No. 20). Washington, D.C.: U.S. Civil Service Commission, Standards Division.

1975 *How to Prepare and Conduct Job Element Examinations*. Personnel Research and Development Center. Washington, D.C.: U.S. Civil Service Commission.

Raju, N.S. 1986 An evaluation of the correlation, covariance, and regression slope models. Paper presented at the meeting of the American Psychological Association, Washington, D.C.

Richardson, Bellows, Henry, and Co. 1983 Technical Reports: The Candidate Profile Record. Washington, D.C.

Rosenfeld, M., B. Shimberg, and R.F. Thornton. 1983 *Job Analysis of Licensed Psychologists in the United States and Canada* . Princeton, N.J.: Center for Occupational and Professional Assessment, Educational Testing Service.

Rosenthal, R. 1984 *Meta-analytic Procedures for Social Research*. Beverly Hills, Calif.: Sage Publications.

Sackett, P.R., N. Schmitt, H. L. Tenopyr, J. Kehoe, and S. Zedeck. 1985 Commentary on "40 questions about validity generalization and meta-analysis." *Personnel Psychology* 38:697-798.

Sackett, P.R., M.M. Harris, and J.M. Orr. 1986 On seeking moderator variables in the meta-analysis of correlation data: a Monte Carlo investigation of statistical power and resistance to type I error. *Journal of Applied Psychology* 71:302-310.

Schmidt, F.L., J.E.Hunter, and K. Pearlman. 1981 Task differences as moderators of aptitude test validity in selection: a red herring. *Journal of Applied Psychology* 66:166-185.

Schmidt, F.L., J.E. Hunter, P.R. Croll, and R.C. McKenzie. 1983 Estimation of employment test validities by expert judgment. *Journal of Applied Psychology* 68:590-601.

Schmidt, F.L., J.E. Hunter. K. Pearlman, and H.R. Hirsh. 1985 Forty questions about validity generalization and meta-analysis. *Personnel Psychology* 38:697-798.

Schneider, B., and N. Schmitt. 1986 *Staffing Organizations*. Glenview, Ill.: Scott-Foresman.

Trattner, M.H. 1982 Synthetic validity and its application to the uniform guidelines validation requirements. *Personnel Psychology* 35:383-397.

Urry, V.W 1978 Some variations on derivation by Primoff and their extensions. Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center.