In A. K. Wigdor and B. F. Green, Jr. (Eds.), Performance Assessment for the Workplace, Vol. 2: Technical Issues. Washington, DC: National Academy of Science Press. 1991.

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. Per-

formance 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

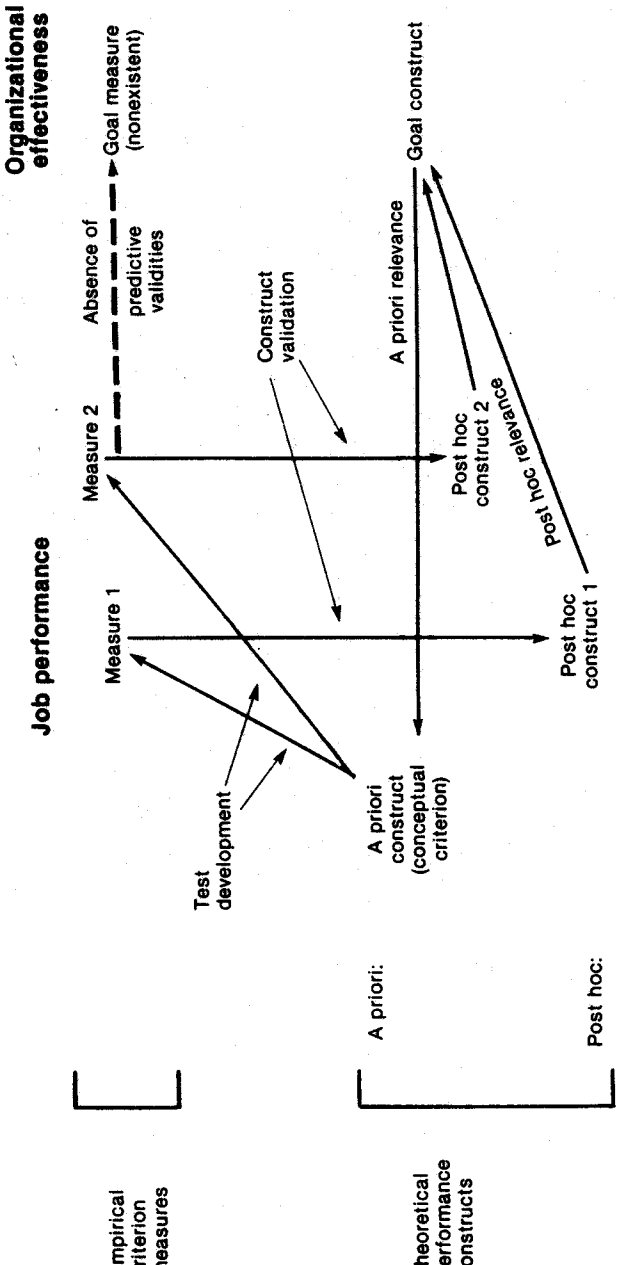

FIGURE 1 A schema summarizing the criterion development and validation process

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 valida-

tion 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

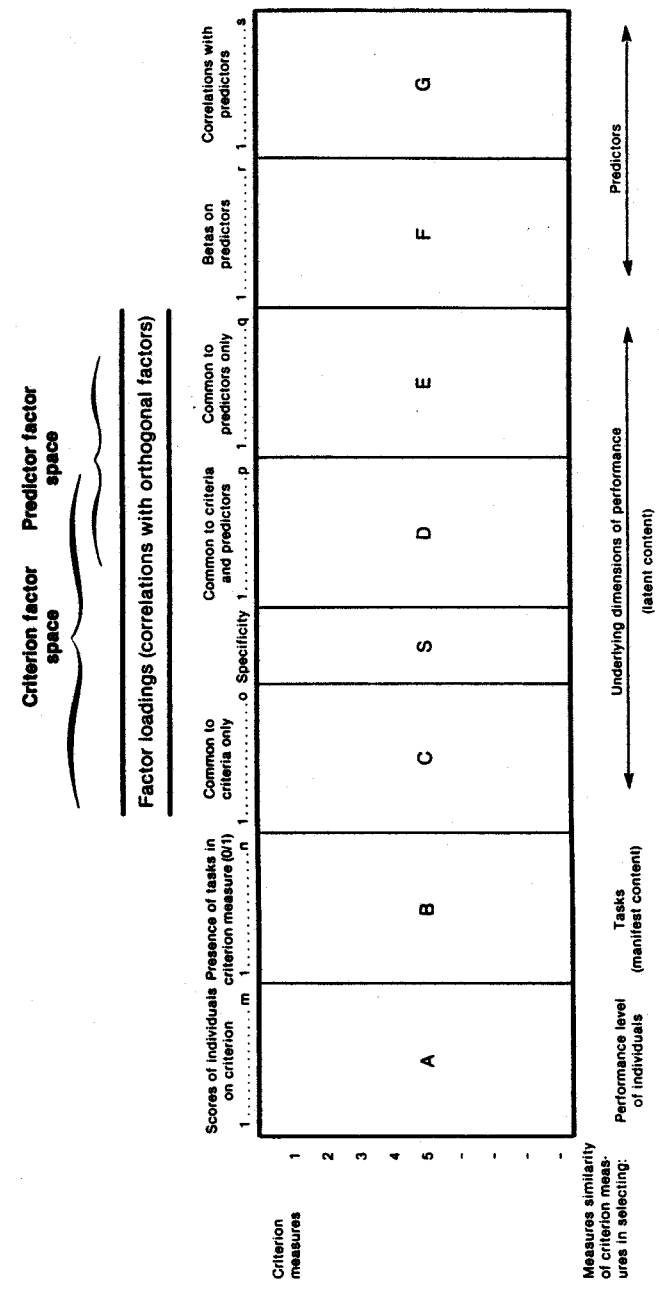

FIGURE 2 The matrices of data (A-G) used for computing alternative estimates of degree of equivalence among different criterion measures

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 pat-

tern 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. *Personnel Psychology* 35:521-551.
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.