EDUCATIONAL MEASUREMENT

Fourth Edition

Sponsored Jointly by
National Council on Measurement in Education and
American Council on Education

Edited by
Robert L. Brennan

AMERICAN COUNCIL ON EDUCATION
PRAEGER
Series on Higher Education

2

Validation

Michael T. Kane
National Conference of Bar Examiners

Measurement uses limited samples of observations to draw general and abstract conclusions about persons or other units (e.g., classes, schools). To validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the claims being made, and this in turn requires a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses. Ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified.

Measurement procedures tend to control irrelevant sources of variability by standardizing the tasks to be performed, the conditions under which they are performed, and the criteria used to interpret the results. Such standardization can be effective in controlling irrelevant variability, but it also restricts the range of observations included in the measurements relative to those that are potentially relevant to the proposed interpretations and uses. The use of test scores to make decisions brings in assumptions about consequences, and these assumptions also merit scrutiny. The challenge is to make the connection between limited samples of observations and the proposed interpretations and uses.

The term "validation" and to a lesser extent the term "validity" tend to have two distinct but closely related usages in discussions of measurement. In the first usage, "validation" involves the development of evidence to support the proposed interpretations and uses; in this usage, "to validate an interpretation or use" is to show that it is justified. In the second usage, "validation" is associated with an evaluation of the extent to which the proposed interpretations and uses are plausible and appropriate. In this sense, "to validate an interpretation or use" is to evaluate its overall plausibility. The first usage implies an advocacy role, in the building of a case for the validity of a proposed interpretation. The second usage implies a more-or-less objective appraisal of the evidence, pro and con.

In this chapter, I will focus on validation as the process of evaluating the plausibility of proposed interpretations and uses, and on validity as the extent to which the evidence supports or refutes the proposed interpretations and uses. However, in the early stages of any validation effort, as the measurement procedure is developed, the more confirmationist usage reflects reality. The test developer is expected to make a case for the validity of the proposed interpretations and uses, and it is appropriate to talk about their efforts "to validate" the claims being made. A mature testing program is expected to stand up to criticism, and the accumulated evidence is to be evaluated in an evenhanded way.

The bulk of this chapter can be divided into three parts. The first part, with two sections, will provide a general conceptual analysis of validation. Section 1 examines the historical development of our conceptions of validity. Section 2 develops an argument-based validation framework, which assumes that the proposed interpretations and uses will be explicitly stated as an argument, or network of inferences and supporting assumptions, leading from observations to the conclusions and decisions. Validation involves an appraisal of the coherence of this argument and of the plausibility of its inferences and assumptions.

The second part, Sections 3 to 6, examines a range of common interpretations and uses and focuses on the kinds of evidence needed to evaluate the inferences and assumptions entailed by these interpretations and uses. These sections cover observable attributes and traits (Section 3), theory-based interpretations (Section 4), qualitative interpretations (Section 5), and decision procedures (Section 6).

The last part, Sections 7 and 8, draws some general conclusions about validation. Section 7, on fallacies, examines some ways in which interpretations and decisions based on test results can go awry, and Section 8 presents some concluding remarks.

Writing a chapter like this involves a number of choices about what to include and what to leave out. Validity now has a long history and a very broad scope. Interest in some issues that seemed to be of central concern in the past (e.g., nomological networks) has faded a bit. Other issues that were given little attention (e.g., social consequences) or taken for granted (standard setting) are receiving more attention. Some aspects of validity are quite technical (e.g., analyses of model fit); some are more philosophical (elucidating the meaning of a score); and some raise broad social issues (fairness, unintended consequences).

I will focus mainly on general strategies for validating the interpretations and uses of measurements. I will not give much attention to technical issues, most of which are discussed in detail in other parts of this volume, but will

attend to the roles of such issues in validation, particularly where the issues come up and how they get resolved.

Validity theory addresses fundamental questions about the meaning and uses of measurements, and therefore, tends to be quite abstract. It seems to have been more successful in developing general frameworks for analysis than in providing clear guidance on how to validate specific interpretations and uses of measurements. This chapter tries to provide a pragmatic approach to validation, involving the specification of proposed interpretations and uses, the development of a measurement procedure that is consistent with this proposal, and a critical evaluation of the coherence of the proposal and the plausibility of its inferences and assumptions.

1. EVOLVING CONCEPTIONS OF VALIDATION

Validation focuses on interpretations, or meanings, and on decisions, which reflect values and consequences. Neither meanings nor values are easily reduced to formulas, literally or figuratively, and some familiarity with how various models of validity developed may be helpful in understanding the models and how they are used (and sometimes abused). An historical approach is particularly useful in understanding the central concept of construct validity, which has undergone several transformations since its introduction about fifty years ago. As a result of these shifts in interpretation, construct validity has accumulated several layers of meaning that are easily blurred.

1.1. Development of the Criterion Model

Between 1920 and 1950, criterion validity came to be seen as the gold standard for validity (Angoff, 1988; Cronbach, 1971; Moss, 1992; Shepard, 1993). In the first edition of *Educational Measurement*, Cureton (1951) defined validity in terms of "the correlation between the actual test scores and the 'true' criterion score" (p. 623), the test-criterion correlation corrected for unreliability in the criterion. Validation was to address the question of how well a test estimates the criterion, which could be defined in terms of "performances of the actual task" (p. 623). A test was considered valid for any criterion for which it provided accurate estimates (Gulliksen, 1950a). Under the criterion model, the variable of interest was assumed to have a definite value for each person, and the goal was to estimate this value as accurately as possible (Ebel, 1961). Given this goal, it was natural to conceive of validity as the correspondence between test scores and criterion scores.

The criterion model came in two versions, concurrent and predictive. *Concurrent validity* studies employed criterion scores obtained at about the same time as the test scores and could be used to validate a proxy measure that would be cheaper, easier, or safer than the criterion. *Predictive validity* studies employed a criterion of future performance (e.g., on the job, in college), which was not available at the time of testing.

The criterion model worked particularly well in those cases where a plausible criterion was readily available. For example, if the test were used to predict some future performance (e.g., in flight training, in a college course, or on the job), evaluations of actual performance could be used as the criterion. Where a good criterion was available, the criterion model provided a simple, elegant, and effective approach to validation, one that could take advantage of sophisticated quantitative methods (Cronbach & Gleser, 1965; Cureton, 1951). For admissions, placement, and employment testing, the criterion model is still the preferred approach (Guion, 1998).

The data from some early criterion-related studies of employment tests seemed to indicate that criterion-related validity coefficients varied substantially, across similar jobs in similar settings. This variability was attributed to differences in the requirements associated with specific jobs in different settings and to differences in the settings (Ghiselli, 1966). The apparent situational specificity of criterion-related data for employment tests led to suggestions that employment tests be validated separately in every situation in which they were to be used. Schmidt and Hunter (1977) examined the variability in criterion-related validity coefficients for employment tests and found that much of the observed variability could be attributed to such statistical artifacts as sampling error, differences in criterion and test reliabilities, and differences in range restrictions; they concluded that validity coefficients could be generalized across settings.

The criterion model can be implemented more-or-less mechanically once the criterion has been defined, but the specification of the criterion typically involves value judgments and a consideration of consequences (Cronbach & Gleser, 1965). Decision procedures based on the criterion model are generally intended "to optimize some later 'criterion' performance" (Cronbach, 1971, p. 445) and therefore involve judgments about the value of the proposed criterion.

The criterion model has two major advantages. First, in many applications, criterion-related evidence is clearly relevant to the plausibility of the proposed interpretations and uses. If the proposed interpretation claims that applicants with higher scores on the test can be expected to exhibit better performance in some activity (e.g., on the job), it would certainly be reasonable to check on this prediction.

Second, criterion-related validity appears to be (and to some extent actually is) objective. Once the criterion is specified, and data on some sample of individuals is collected, a validity coefficient can be computed in a straightforward way. Of course, the choice of a criterion and the selection of individuals to be included in the study involve a number of value judgments, but once these choices are made and the data are collected, the analyses can be straightforward.

1.1.1. Limitations of the Criterion Model

The main limitation in the criterion model was and is the difficulty in obtaining an adequate criterion. In some cases (e.g., achievement tests), it may be difficult to implement a criterion that is clearly better than the test itself, and in other cases (e.g., intelligence, creativity), it may be difficult to even conceptualize a satisfactory criterion (Cronbach, 1971, 1980a, b; Guion, 1998; Lord & Novick, 1968). As Ebel suggested:

> The ease with which test developers can be induced to accept as criterion measures quantitative data having the slightest appearance of relevance to the trait being measured is one of the scandals of psychometry. (1961, p. 642)

Once one begins to question some criteria, it becomes clear that all criteria are questionable.

In addition to the many practical problems that plague criterion-related validity studies, the criterion model faces a fundamental problem. How can the criterion be validated? Even if a second criterion can be identified as a basis for validating the initial criterion, this simply pushes the problem back one step (Ebel, 1961). Without some other way to validate some criterion measures, we clearly face either infinite regress or circularity. The criterion model is quite useful in validating secondary measures of an attribute, assuming that some primary measure is available to be used as a criterion, but, ultimately, it cannot be used to validate the criterion. At some point, the criterion has to be validated in another way.

1.2. Content Model

How can a criterion be validated without appealing to another criterion? One option would be to validate the criterion by establishing a rational link between the procedures used to generate the criterion scores and the proposed interpretation or use of the scores (Cureton, 1951; Ebel, 1961; Gulliksen, 1950b; Rulon, 1946). Where a sample of some type of performance (e.g., typing, flying an airplane, playing the piano) is used to draw conclusions about level of skill in that kind of performance, a good case for the validity of the proposed interpretation can be made on rational grounds (Cronbach, 1971; Cureton, 1951; Ebel, 1961; Kane, Crooks, & Cohen, 1999).

The content model interprets test scores based on a sample of performances in some area of activity as an estimate of overall level of skill in that activity. Assuming that a person's performance is evaluated on a sample of tasks from a domain, it is legitimate to take the observed performance as an estimate of overall performance in the domain, if (a) the observed performances can be considered a representative sample from the domain, (b) the performances are evaluated appropriately and fairly, and (c) the sample is large enough to control sampling error (Guion, 1977).

Validity claims for assessments based on samples of the performance of interest can be challenged on various grounds (e.g., by suggesting that the samples are biased), but such sampling models do provide a basis for validation without resorting to external criteria (Cronbach, 1971; Kane, 1982). This is especially true if the relevant content domain has been defined with care, the tasks have been sampled (or developed) in a way that makes them representative of the domain, the observations were made using procedures that would tend to control random and systematic errors, and the performances were evaluated appropriately. We can find out how well somebody can perform a task by evaluating samples of their performance on the task. This approach tends to work especially well for tests of specific skills, but it can also be applied to more broadly defined measures of achievement (e.g., Flockton & Crooks, 2002).

The content model has most frequently been applied measures of academic achievement. A content domain is outlined in the form of a test plan or blueprint, which may involve several dimensions (e.g., content per se, cognitive level, item type), with different numbers of items assigned to each cell in the plan. The items are not sampled from the domain; they are created to match the test specifications (Loevinger, 1957), and to the extent that they do, they may be considered to be representative of the content domain described by the test plan.

1.2.1. Limitations of the Content Model

The content model has been subject to a number of criticisms. In particular, content-related validity evidence tends to be subjective and to have a confirmatory bias. Content-based analyses tend to rely on judgments about the relevance and representativeness of test tasks. When these judgments are produced by test developers, they have a natural tendency to confirm the proposed interpretation.

The content model is especially problematic when it is used to argue for the validity of claims about cognitive processes or other theoretical constructs:

> Judgments about content validity should be restricted to the operational, externally observable side of testing. Judgments about the subject's internal processes state hypotheses, and these require empirical *construct* validation. (Cronbach, 1971, p. 452; italics in original)

Messick (1989) argued that content-based validity evidence does not involve test scores or the performances on which the scores are based and therefore cannot be used to justify conclusions about the interpretation of test scores. Messick (1989) described content-validity evidence as providing support for "the domain relevance and representativeness of the test instrument" (p. 17) but saw it as playing a limited role in validation because it doesn't provide direct evidence for the "inferences to be made from test scores" (p. 17).

Content-related evidence has an important role to play in validation, but in most cases, it is a limited role. Evidence for the representativeness of the test tasks and the generalizability of scores can make a basic positive case for the validity of an interpretation in terms of expected performance over some universe of possible performances, but to go beyond that basic interpretation, other kinds of evidence are needed.

1.3. The Construct Model

By the early 1950's, the criterion model was well developed, and the content model was used to establish the plausibility of criterion measures. The content model also provided the main framework for validating measures of achievement and measures of various dispositions. However, models for the validation of measures of theoretical attributes were lacking.

In the early 1950s, the APA Committee on Psychological Tests sought to identify the kinds of evidence needed to justify the psychological interpretations that were "the stock-in-trade of counselors and clinicians" (Cronbach,

1989, p. 148). The resulting proposals for construct validation were incorporated in the *Technical Recommendations* (American Psychological Association, 1954), and were further developed by Cronbach and Meehl (1955). In 1971, Cronbach reviewed the events leading up to the 1955 paper:

> The rationale for construct validation (Cronbach & Meehl, 1955) developed out of personality testing. For a measure of, for example, ego strength, there is no uniquely pertinent criterion to predict, nor is there a domain of content to sample. Rather, there is a theory that sketches out the presumed nature of the trait. If the test score is a valid manifestation of ego strength, so conceived, its relations to other variables conform to the theoretical expectations. (Cronbach, 1971, pp. 462–463)

Cronbach and Meehl (1955) framed their discussion of construct validity in terms of the hypothetico-deductive (HD) model of scientific theories, in which a theory consists of a network of relationships linking theoretical constructs to each other and to observable attributes. Each theoretical construct is implicitly defined by its role in the theory.

Under the HD model of theories (Suppe, 1977), if the predictions derived from the theory do not agree with observations, then either the theory is wrong, the measurements are not appropriate indicators of the constructs in the theory, or some ancillary assumption was violated. If diverse predictions based on the theory are confirmed, the theory and the interpretation of scores in terms of the theory are supported.

This model assumes the existence of a well-defined theory from which empirical predictions can be derived. Cronbach and Meehl (1955) recognized the limitation inherent in this requirement:

> The idealized picture is one of a tidy set of postulates which jointly entail the desired theorems ... In practice, of course, even the most advanced physical sciences only approximate this ideal ... Psychology works with crude, half-explicit formulations. (pp. 293–294)

Nevertheless, they suggested that "the network still gives the constructs whatever meaning they do have" (p. 294). The construct is defined by its role in the theory.

Soon after Cronbach and Meehl's (1955) exposition of construct validity, Loevinger (1957) published a seminal analysis of the relationship between tests and psychological theory. Where Cronbach and Meehl basically started with the construct and its theory and posed the question of whether the test was an adequate measure of the construct, Loevinger started from the test and then asked about the plausibility of the proposed interpretation:

> to what extent does the test measure a trait that "really" exists. And how well does the proposed interpretation correspond to what is measured by the test? (Loevinger, 1957, p. 643)

Taking test construction as the starting point, Loevinger (1957) partitioned construct validity into a substantive component (which focused on a theory-based analysis of test content), a structural component (which focused on the internal structure of the test in relation to that of the target construct), and an external component (which focused on relationships to other test and non-test variables and on potential sources of systematic error). Loevinger (1957) considered the criterion and content models to be *ad hoc* because they depended on specific contingencies, proposing that "construct validity is the whole of the subject from a systematic, scientific point of view" (p. 461).

1.4. Evolution of the Construct Model 1955–1989

Cronbach and Meehl (1955) presented construct validity as an alternative to the criterion and content models; it was to be used "whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined" (1955, p. 282), and "for which there is no adequate criterion" (1955, p. 299). Shepard (1993, p. 416) has suggested that Cronbach and Meehl introduced construct validity as "the weak sister" to be used when a real criterion is not available.

However, Cronbach and Meehl (1955, p. 282) went on to say that "determining what psychological constructs account for test performance is desirable for almost any test." That is, even if the test is initially validated using criterion or content evidence, the development of a deeper understanding of the constructs or processes accounting for test performance requires a consideration of construct-related evidence. So, Cronbach and Meehl (1955) suggested that construct validity was a fundamental concern, but they did not present it as a general organizing framework for validity.

The 1966 Standards distinguished construct validity from other approaches to validity, particularly criterion-related validity, by suggesting that "construct validity is relevant when the tester accepts no existing measure as a definitive criterion" (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education [APA, AERA, & NCME], 1966, p. 13). The 1974 *Standards* (APA et al., 1974) continued along this track, listing four kinds of validity associated with four kinds of interpretation (predictive and concurrent validities, content validity, and construct validity).

Cronbach (1971) distinguished several approaches to validation and associated construct validation with theoretical variables for which "there is no uniquely pertinent criterion to predict, nor is there a domain of content to sample" (p. 462) and said that any description "that refers to the person's internal processes (anxiety, insight) invariably requires construct validation" (p. 451). He also emphasized the need for an overall evaluation of validity involving multiple kinds of evidence:

> Validation of an instrument calls for an integration of many types of evidence. The varieties of investigation are not alternatives any one of which would be adequate. The investigations supplement one another ... For purposes of exposition, it is necessary to subdivide *what in the end must be a comprehensive, integrated evaluation of the test.* (Cronbach, 1971, p. 445; italics in original)

The construct model continued to serve as one of several possible approaches to validation, but Cronbach also emphasized the need to integrate different kinds of validity evidence in evaluating the proposed interpretations and uses of test scores.

The need to specify and then evaluate the proposed interpretation was also apparent in the development of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), which deals with the precision (or reliability) of measurements. Generalizability theory, which provides analyses of error variance, required the test developer to specify the conditions of observation (e.g., content areas, test items, raters, occasions) over which generalization is to occur and then to evaluate the impact of the sampling of different kinds of conditions of observation on observed scores.

By the late 1970s, two opposing trends were evident in the development of validity theory. On the one hand, there had been a longstanding interest in a clear specification of the kinds of evidence needed to validate particular interpretations and uses of test scores. On the other hand, there was a perceived need to develop a unified conception of validity. The 1985 Standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1985) sought to resolve this tension by treating validity as a unified concept while recognizing that different kinds of evidence were relevant to different kinds of interpretations.

The need for specific rules for validation was particularly acute in employment testing, where employers wanted to know what they had to do to satisfy legal requirements for fairness in testing. For practical reasons, employers tended to prefer content-based strategies. Predictive evidence clearly made sense in employment testing but was expensive and often not feasible. The Uniform Guidelines (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice [EEOC, CSC, DoL, & DoJ], 1978) developed by federal agencies for the implementation of civil rights legislation required that evidence for validity be provided if a test had adverse impact on any protected group. The Uniform Guidelines accepted content or construct evidence but expressed a clear preference for criterion-related, predictive evidence (EEOC et al., 1978; Landy, 1986).

Nevertheless, validity theorists (Cronbach, 1980b; Guion, 1977, 1980; Messick, 1975, 1981; Tenopyr, 1977) preferred a more unified approach and expressed concern about the growing tendency to treat validation methodology as a toolkit, with different models to be employed for different assessments. The criterion model would be used to validate selection and placement decisions, the content model would be used to validate achievement tests, and the construct model would be used for theory-based explanations.

1.4.1. Construct Validation as the Basis for a Unified Model of Validity

Loevinger's (1957, p. 636) suggestion that "since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" took a while to catch on, but by the early 1980s, the construct model was widely accepted as a general approach to validity (Anastasi, 1986; Embretson, 1983; Guion, 1977; Messick, 1980, 1988, 1989). This broad conception of construct validity was taken to encompass all evidence for validity, including content and criterion evidence, reliability, and the wide range of methods associated with theory testing. Messick (1988, 1989) adopted a broadly defined version of the construct model as a unifying framework for validity. He relegated the content model to a subsidiary role in supporting the relevance of test tasks to the constructs of interest, and he treated the criterion model as an ancillary methodology for validating secondary measures of a construct against its primary measures.

The adoption of the construct model as the unified framework for validity had three major positive effects. First, the construct model tended to focus attention on a broad array of issues inherent in the interpretations and uses of test scores, and not simply on the correlation of test scores with specific criteria in particular settings and populations. Second, it emphasized the pervasive role of assumptions in score interpretations and the need to check these assumptions. Finally, the construct model allowed for the possibility of alternative interpretations and uses of test scores.

In the third edition of *Educational Measurement*, Messick (1989, p. 13) defined validity as:

> an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. [italics in original]

He summarized his framework for validity evidence in terms of two interconnected "facets" of validity: the justification for testing (an evidential basis or a consequential basis) and the function or outcome of testing (interpretation or use), yielding a two-by-two matrix. Messick (1989) associated the evidential basis for score interpretation with construct validity and the evidential basis for test use with construct validity and relevance or utility. That is, all interpretations of test scores are to be supported by construct validation, and the appropriateness of the scores for a particular purpose is to be evaluated in terms of the relevance of the construct to the purpose at hand. Under this model, the evaluation of test use is a two-step process, from score to construct and from construct to use.

Messick gave the consequential basis for validity equal billing with the evidential basis. According to Messick (1989), the consequential basis of test interpretation depends on the value implications of the construct label and on the assumptions (theoretical and ideological) supporting the interpretation. Messick associated the consequential basis of test use with "the appraisal of both potential and actual social consequences of the applied testing" (Messick, 1989, p. 20).

1.4.2. Criteria for the Adequacy of Validation Efforts

The broad, unified version of construct validity was quite appealing as a conceptual framework for validity but it did not provide clear guidance for the validation of a test-score interpretation or use. In the absence of strong theories, construct validity tends to be very open ended, and it is not clear where to begin or how to gauge progress. Cronbach suggested that construct validation be viewed "as an ever-

extending inquiry into the processes that produce a high or low test score and into the other effects of those processes" (Cronbach, 1971, p. 452), and subsequently characterized validation as "a lengthy, even endless process" (Cronbach, 1989, p. 151). According to Anastasi (1986), "almost any information gathered in the process of developing or using a test is relevant to its validity" (p. 3). If all data are relevant to validity, where should one start, and how much evidence is needed to adequately support a validity claim?

The call for serious consideration of alternative interpretations within the construct model offered one basis for choosing studies to conduct. Those studies that address the strongest competing interpretations would be given the highest priority. However, in practice, most validation research is conducted by test developers and tends to have a confirmationist bias. As Cronbach (1989) observed, "Falsification, obviously, is something we prefer to do unto the constructions of others" (p. 153). In itself, the unified model of construct validity does not necessarily provide clear guidelines for evaluating the adequacy of validation efforts.

1.5. General Principles Emerging from the Construct-Validation Model

Over the period from 1955 to 1989, at least three aspects of the construct-based model gradually emerged as general principles of validation that transcended the theory-dependent context in which construct validity was introduced.

First, by focusing on the role of potentially complex theories in specifying what attributes mean, Cronbach and Meehl (1955) increased awareness of the need to specify the proposed interpretation before evaluating its validity. Within the criterion model, it is relatively easy to develop validity evidence based on a test-criterion correlation without examining the rationale for the criterion too carefully. In marked contrast, the development of construct-related validity evidence requires that the proposed construct interpretation be elaborated in some detail. As a result, between 1955 and 1989, the emphasis gradually shifted from the validation of the test (as a measure of an existing criterion) to the development and validation of a proposed interpretation of test scores (Cronbach, 1971, 1982; Loevinger, 1957).

Second, Cronbach and Meehl (1955) recognized that construct validation would involve an extended research program. The validation of a measure of a theoretical construct involves the specification of a theory and of measures of the constructs in the theory and the empirical evaluation of predictions derived from the theory. The construct-validity model requires a research program, rather than a single empirical study (Cronbach, 1971).

Third, construct validity's focus on theory testing led to a growing awareness of the need to challenge proposed interpretations and to consider competing interpretations. Cronbach and Meehl (1955) did not give much direct attention to the evaluation of alternate interpretations, but this notion is implicit in their focus on theory and theory testing, and it was made fully explicit in subsequent work on construct validity (Cronbach, 1971, 1980a, b, 1982; Embretson, 1983; Messick, 1989).

By the mid-1980's, the model introduced by Cronbach and Meehl (1955) had developed into a general methodology for validation, as these three methodological principles (the need for an explicit statement of the proposed interpretation, the need for extended analysis in validation, and the need to consider alternate interpretations) were accepted as basic principles of validation.

1.6. Argument-Based Approach to Validity

Cronbach (1988) proposed that the validation of score interpretations and uses be based on the logic of evaluation argument (Cronbach, 1982; House, 1980). In program evaluation, it is clearly necessary to specify the program to be evaluated and the contexts in which it will be implemented, and any satisfactory evaluation involves a program of research, rather than a single empirical study. In program evaluation, alternative explanations for any observed changes need to be ruled out through experimental controls and/or through an explicit evaluation of alternative explanations. That is, validation and program evaluation face many of the same issues.

Cronbach suggested that the *validity argument* is to provide an overall evaluation of the intended interpretations and uses of test scores by generating a coherent analysis of all of the evidence for and against the proposed interpretation/use, and to the extent possible, the evidence relevant to plausible alternate interpretations and decision procedures (Cronbach, 1988). Similarly, Messick's (1989) definition of validity required an evaluative judgment of "the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores" (p. 12).

In order to evaluate a proposed interpretation of test scores, it is necessary to have a clear and fairly complete statement of the claims included in the interpretation and the goals of any proposed test uses. The proposed interpretations and uses can be specified in detail by laying out the network of inferences and assumptions leading from the test performances to the conclusions to be drawn and to any decisions based on these conclusions (Crooks, Kane, & Cohen, 1996; Kane, 1992; Shepard, 1993). The most recent edition of the *Standards* (AERA et al., 1999) has adopted a similar approach, suggesting that, "Validity logically begins with an explicit statement of the proposed interpretation of test scores along with a rationale for the relevance of the interpretation to the proposed use" (p. 9).

The general approach is consistent across applications. The inferences included in the interpretations and uses are to be specified, these inferences and their supporting assumptions are to be evaluated using appropriate evidence, and plausible alternative interpretations are to be considered. However, the model is responsive to differences in proposed interpretations and uses and to the context in which the scores are to be used. The specific inferences and assumptions tend to change from one application to another, and therefore the evidence required to support these claims tends to change, but validation always involves the specification (the interpretive argument) and evaluation (the validity argument) of the proposed interpretations and uses of the scores.

The argument-based approach to validity reflects the general principles inherent in construct validity without an emphasis on formal theories. The interpretive argument is to provide a clear statement of the inferences and assumptions inherent in the proposed interpretations and uses of test results, and these inferences and assumptions are to be evaluated in a series of analyses and empirical studies. Individual studies in a validity argument may focus on statistical analyses, content analyses, or relationships to criteria, but the validity argument as a whole requires the integration of different kinds of evidence from different sources. Plausible rival interpretations can provide particularly effective challenges to a proposed interpretation.

The main advantage of the argument-based approach to validation is the guidance it provides in allocating research effort and in gauging progress in the validation effort. The kinds of validity evidence that are most relevant are those that support the main inferences and assumptions in the interpretive argument, particularly those that are most problematic. In this vein, the 1999 *Standards* suggested that: "validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use" (AERA et al., 1999, p. 9). If some inferences in the argument are found to be implausible given the evidence, the interpretive argument needs to be either revised or abandoned. The proposed interpretations and uses of the test scores determine the kinds of evidence that are needed for validation.

2. VALIDITY AS ARGUMENT

To validate a proposed interpretation or use of test scores is to evaluate the rationale for this interpretation or use. The evidence needed for validation necessarily depends on the claims being made. Therefore, validation requires a clear statement of the proposed interpretations and uses.

2.1. Interpretive Arguments and Validity Arguments

Validation employs two kinds of argument. An *interpretive argument* specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances.

The *validity argument* provides an evaluation of the interpretive argument (Cronbach, 1988). To claim that a proposed interpretation or use is valid is to claim that the interpretive argument is coherent, that its inferences are reasonable, and that its assumptions are plausible. The analysis "should make clear, and to the extent possible, persuasive the construction of reality and the value weightings implicit in a test and its application" (Cronbach, 1988, p. 5).

The interpretive argument involves inferences leading from observed performances to the claims based on these performances. Each inference involves an extension of the interpretation or a decision. The inferences take the form of "if-then" rules (e.g., if the observed performance has certain characteristics, then the observed score should have a certain value, if an applicant's score is above some cutscore, the applicant is admitted). The assumptions supporting each inference are those that would provide adequate backing for the inference if they were accepted as true; for example, the assumption that a sample of performance is representative of some universe of possible performances supports an inference from the mean score for the sample to the mean over the universe.

There are, potentially, a large number of assumptions in any interpretive argument. We take many of these assumptions for granted, at least until evidence to the contrary develops. On written tests, we typically assume that students can read the questions and understand the instructions, unless special circumstances suggest otherwise. But some of the assumptions cannot be taken for granted, even in ordinary cases, and therefore need to be evaluated. For example, the adequacy with which an achievement test covers a content domain is almost always questionable, because the development of such tests involves a very large number of choices about what to include and what to leave out, and any or all of these choices may be questioned.

The interpretive argument makes the reasoning inherent in the proposed interpretations and uses explicit so that it can be better understood and evaluated. By outlining the claims to be evaluated, it provides a framework for validation. For example, if the interpretive argument includes a statistical generalization from observed performance on a sample of occasions to expected performance over a universe of possible occasions, the validity argument should evaluate the dependability of this generalization over occasions. If the interpretation does not assume generalization over occasions, empirical evidence for consistency over occasions would be irrelevant to the validity of the proposed interpretation and might even cast doubt on its validity. If the attribute being measured is expected to change rapidly over some period of time (e.g., level of achievement during a period of intensive instruction), invariance of the scores over time may indicate a lack of sensitivity in the measurement procedure. Evidence for generalizability over a facet (e.g., items, rater, conditions of measurement) is to be included in the validity argument if and only if the interpretive argument involves generalization over that facet.

Similarly, if the scores are used to predict future performance, the relationship between scores and measures of future performance needs to be examined; on the other hand, if the scores are used to certify current competence, predictive accuracy is not necessarily a concern, but the appropriateness of the passing score becomes a salient issue. If the interpretation depends on a scaling model (e.g., for scoring responses) the fit of the model to the data needs to be examined. If the interpretation does not depend on a scaling model, there is no need to evaluate how well the data fit that model. It is the plausibility of the proposed interpretations and uses that is to be evaluated.

2.2. An Example—Placement Testing

Placement tests are widely used in higher education to assign students within a sequence of related courses. It is assumed that the competencies developed in earlier courses serve as prerequisites for later courses (van der Linden,

1998). The goal is to assign students to courses that will be optimal for them in some sense (i.e., to a course that will be demanding but not overwhelming for the student). The ideal course would be one focusing on competencies that the student has not mastered, but for which the student has mastered all prerequisites.

The development of placement tests usually begins with the identification of the competencies developed in the courses. Once the relevant competencies have been identified, a test covering these competencies can be developed. The intent is not to develop a good, general measure of achievement in the area, but rather to focus on the specific competencies that are needed for courses in the sequence. The intended uses of the placement system drive the test-development process.

The rationale for placement testing is essentially a special case of what Cronbach and Snow (1977) called an aptitude-treatment interaction, with current level of skill in the competencies defining the aptitude, and the different courses constituting different treatments. A student with a very low score on the placement test, indicating a low level of skill in the competencies, would be expected to do poorly in any course beyond the first, which has few if any prerequisites. Students with higher scores would be expected to profit most from higher level courses.

The decision procedures for test-based placement systems are usually defined by specifying a set of cutscores on the test-score scale. Students with scores below the lowest cutscore are placed in the lowest level course; students with scores between the first and second cutscores are placed in the second course, and so on.

For placement tests, the cutscores are generally set empirically using the data from studies in which students who are currently completing the courses take the placement test, and their scores are matched with their grades in the course. Assuming that it is desirable to place each student in the highest level course in which he or she has a good chance of success, and that a student with a grade of B or better in one course is likely to be successful in the next course in the sequence, the lower cutscore for assignment to a course can be set to correspond to an expected score of B in the next lower course. Note that the data used to set the cutscore also provide empirical support for the cutscore and for the relationship between test scores and performance in the different courses.

2.2.1. *Interpretive Argument for Placement Testing*

An interpretive argument for placement systems, which is outlined in Table 2.1, includes four major inferences: scoring, generalization, extrapolation, and a decision. Each of these inferences depends on several assumptions. The interpretive argument may also involve various technical inferences and assumptions (e.g., equating, scaling) that are not discussed here.

The scoring inference employs a scoring rule to assign a score to each student's performance on the test tasks. For multiple-choice tests, the scoring rule consists of an answer key for the test. For performance assessments, the scoring rule, or rubric, provides guidelines for grading student performances. The scoring inference relies on two basic assumptions: that the scoring criteria are reasonable and that they are applied appropriately.

TABLE 2.1 Interpretive Argument for a Placement Testing System

I1: ***Scoring***: from observed performance to an observed score
- A1: The scoring rule is appropriate.
- A2: The scoring rule is applied accurately and consistently.

I2: ***Generalization***: from observed score to universe score
- A1: The observations made in testing are representative of the universe of observations defining the testing procedure.
- A2: The sample of observations is large enough to control sampling error.

I3: ***Extrapolation***: from universe score to the level of skill
- A1: The test tasks require the competencies developed in the courses and required in subsequent courses.
- A2: There are no skill irrelevant sources of variability that would seriously bias the interpretation of scores as measures of level of skill in the competencies.

I4: ***Decision***: from conclusion about level of skill to placement in a specific course.
- A1: Performance in courses, beyond the initial course, depends on level of skill in the competencies developed in earlier courses in the sequence.
- A2: Students with a low level of skill in the prerequisites for a course are not likely to succeed in the course.
- A3: Students with a high level of skill in the competencies taught in a particular course would not benefit much from taking the course.

The generalization inference extends the interpretation from the observed score to a claim about expected performance over a larger universe of observations allowed by the testing procedure. Generalization depends on two assumptions: that the sample of observations is representative of the universe and that the sample is large enough to control sampling error. If both of these assumptions are met, sampling theory provides support for generalization from the observed score to the expected score over the universe.

The extrapolation inference extends the interpretation from test performance to a claim about competencies required in the courses, and therefore, to expectations about how well the student is likely to do in various courses. The extrapolation inference assumes that the test tasks provide adequate measures of the competencies of interest (those developed in the courses) and are not overly influenced by extraneous factors (e.g., test format).

The decision inference assigns students to courses based on their test scores. This inference depends on a number of value assumptions. It assumes that it is desirable that students be placed in the highest level courses in which they are likely to do fairly well, and that it is important that students not do badly in the courses in which they are placed. These two assumptions are value judgments about possible outcomes of instruction; they represent widely held values, but they are value judgments. As Cronbach (1971) pointed out, any evaluation of a placement decision rests on the assumption that, "the outcome will be

more satisfactory under one course of action than another" (p. 448).

2.2.2. Validity Argument for Placement Tests

The validity argument provides an evaluation of the interpretive argument and would begin with a review of the argument as a whole to determine if it makes sense. Assuming that the interpretive argument (e.g., that in Table 2.1) is considered reasonable, its inferences and assumptions would be evaluated using appropriate evidence.

Scoring, generalization, and extrapolation are examined in some detail in Section 3 of this chapter, and decision inferences are analyzed in Section 6. Therefore, the validity argument for placement tests will only be sketched here. Support for scoring inferences relies mainly on expert judgment about the appropriateness of the scoring rule and the thoroughness of quality control procedures. Statistical analyses of the agreement among raters are called for if the scoring involves ratings. If scaling or equating models are employed in generating the final scores, the fit of these models to the data would also be checked empirically.

The analysis of generalizability involves reliability studies or generalizability studies, as well as judgments about the representativeness of the samples of observations included in the test.

Evaluations of the extrapolation inference may be based on judgments about the overlap between the skills measured by the test and those needed in the courses and/or by empirical analyses of the relationship between test scores and measures of performance in courses (e.g., course grades).

Finally, the evaluation of the placement decisions would involve an analysis of the positive and negative consequence resulting from the decisions, relative to those for alternative decision procedures.

In general, several different kinds of evidence would be relevant to each inference and its supporting assumptions, including expert judgment, empirical studies, the results of previous research, and value judgments.

2.3. The Interpretive Argument as Mini-Theory

The interpretive argument is analogous to a scientific theory (Suppe, 1977). Just as a scientific theory provides a general framework for interpreting certain observed phenomena and achieving certain goals, the interpretive argument provides a framework for the interpretation and use of the scores on a test. Both scientific theories and interpretive arguments are evaluated in terms of their clarity, their internal consistency, and the plausibility of their inferences and assumptions, some of which are evaluated by checking their implications against empirical data (Popper, 1962). And just as the applicability of a theory may be questioned in a particular case, the applicability of an interpretive argument to a test performance may be questioned in a particular case.

The generic form of the interpretive argument represents the proposed interpretations and uses of test scores. It is applied every time test results are used to draw conclusions or make decisions. It specifies the reasoning involved in getting from the test results to the conclusions and decisions based on these results. A validity argument that yields a positive evaluation of the reasoning in the interpretive argument provides support for the appropriateness of the proposed interpretations and uses of test results. Neither the interpretive argument nor the validity argument has to be developed anew for each person's test performance.

However, even if the interpretive argument works well in most cases, it may fail in situations in which one or more of its assumptions fails to hold. For example, even if the validity of a test as a measure of achievement in a course is well documented, this interpretation may fail in a special case.

2.4. The Development and Appraisal Stages of Validation

The validation of a proposed interpretive argument can be separated into two stages. In the *development stage*, the focus is on the development of the measurement procedure and the corresponding interpretive argument. In the *appraisal stage*, the focus is on the critical evaluation of the interpretive argument.

In practice, these two stages are likely to overlap considerably, but it is useful to emphasize the shift that needs to occur at some point in the process. Initially the goal is to develop an assessment program that supports the proposed interpretations and uses of scores. It is appropriate (and probably inevitable) that the test developers have a confirmationist bias; they are trying to make the testing program as good as it can be. However, at some point, especially for high-stakes testing programs, a shift to a more arms-length and critical stance is necessary in order to provide a convincing evaluation of the proposed interpretations and uses.

2.4.1. The Development Stage: Creating the Test and the Interpretive Argument

The test developers have some interpretations and/or uses in mind when they begin developing a test. This initial statement of the proposed interpretations and/or uses may be quite general (e.g., to place each student into an appropriate course) but some goal is needed to get started.

The developers decide on a general approach to achieving the goal at hand and then develop the measurement procedure and the accompanying interpretive argument. Efforts to make the measurement procedure consistent with the proposed interpretations and uses provide support for the plausibility of the interpretive argument.

In addition, efforts to identify and control possible sources of extraneous variance may help to rule out certain alternative interpretations. For example, if the test is not intended to be speeded, pilot testing can be used to set appropriate time limits and therefore to make an alternative interpretation in terms of speed less likely. The development stage has a legitimate confirmationist bias; its purpose is to develop the test and a plausible interpretive argument that reflects the proposed interpretations and uses of test scores.

A basic iterative strategy for developing the test and the corresponding interpretive argument can be described in terms of three steps. First, an interpretive argument is outlined (e.g., the interpretive argument for placement tests outlined in the last section) and a test plan is developed.

Difficulty in specifying an interpretive argument for a proposed interpretation or use may indicate a fundamental problem. If it is not possible to come up with a test plan and a plausible rationale for a proposed interpretation or use, it is not likely that this interpretation or use will be considered valid.

Second, the test would be developed. Efforts to develop a test that is consistent with the interpretive argument (e.g., by identifying the relevant competencies and by constructing test tasks that measure these competencies) can make a preliminary case for the interpretive argument.

Third, the inferences and assumptions in the interpretive argument would be evaluated to the extent possible during test development. Any weaknesses unveiled by these analyses may indicate the need to modify the interpretive argument or the test. If the needed changes are substantial, the first two steps would be repeated. This process continues until the test developers are satisfied with the fit between the test and the interpretive argument.

This iterative process is much like the process of initial theory development and refinement in science, with the interpretive argument playing the role of a theory. The initial form of the theory is proposed. Any weaknesses identified in the theory are corrected, if possible, by changing some assumptions in the theory or by changing the scope of the theory (i.e., the range of cases to which it applies). If the evidence reveals inconsistencies that can't be resolved, the theory (or the interpretive argument) is rejected, but this is a last resort. The evidence produced during the development stage tends to be confirmationist; if a problem is identified, it is fixed, if it can be fixed.

2.4.2. Appraisal Stage: Challenging the Interpretive Argument

At some point, the development process is considered complete, and it is appropriate for the validation effort to adopt a more neutral or even critical stance:

> a proposition deserves some degree of trust only when it has survived serious attempts to falsify it. (Cronbach, 1980b, p. 103)

During the appraisal stage, the test is taken as a finished product and the overall plausibility of the interpretive argument is examined.

Some of the evidence for this critical evaluation will come from the development stage. It may seem like a low hurdle, but one product that the development stage is expected to deliver is an explicit, coherent interpretive argument linking test performances to the proposed interpretations and uses. If this minimal expectation has not been met, a critical evaluation of the proposed interpretive argument is premature.

For low-stakes applications involving relatively plausible inferences and assumptions, the evidence derived from the development phase may be sufficient for the appraisal of the interpretive argument. For example, the teacher who uses a performance assessment to provide feedback to students is likely to be on safe ground.

For high-stakes applications, a more extensive appraisal is called for, especially if the interpretive argument makes ambitious claims (e.g., about how well test takers will do in a specific course or institution) that are not fully evaluated during test development. Furthermore, the validity argument for high-stakes testing programs has to be persuasive to a number of audiences (Cronbach, 1988). To meet these varied criteria, a critical evaluation of the proposed interpretations and uses is needed.

During the appraisal stage, studies of the most questionable assumptions in the interpretive argument are likely to be most informative, but it is also prudent to check any inferences and assumptions that are easy to check (Cronbach, 1982, 1988). If the proposed interpretive argument can withstand these challenges, confidence in the claims increases. If they do not withstand these challenges, then either the assessment procedure or the interpretive argument has to be revised or abandoned.

Cronbach (1989) proposed four criteria for identifying the empirical studies to be pursued by the test evaluator:

1. Prior uncertainty: Is the issue genuinely in doubt?
2. Information yield: How much uncertainty will remain at the end of a feasible study?
3. Cost: How expensive is the investigation in time and dollars?
4. Leverage: How critical is the information for achieving consensus in the relevant audience? (Cronbach, 1989, p. 165)

Resources are always limited, and choices have to be made.

The appraisal stage also involves a search for hidden assumptions and investigations of alternative possible interpretations of the test scores. The interpretive argument specified during the development phase may be incomplete in various ways. For example, the criteria used for scoring performances may make a number of value judgments that are not spelled out in any detail. A critical appraisal of that part of the interpretive argument would seek to make any such assumptions explicit and to subject them to scrutiny, perhaps by individuals who espouse different values. Similarly, the interpretive argument may include various assumptions about the outcomes that are likely to result from the proposed uses of the test scores and about the relative values of different outcomes; presumably, these assumptions should be made explicit and critically reviewed.

An emphasis on specific types of evidence to evaluate specific assumptions provides some protection against inappropriate interpretations and uses of test scores. To the extent that the interpretive argument is specified in some detail, gaps and inconsistencies are harder to ignore, and a lack of evidence for one or more steps in the argument may be more apparent. An interpretive argument that has survived all reasonable challenges to its coherence and plausibility can be provisionally accepted, with the understanding that new evidence may undermine its credibility in the future.

Without the discipline imposed by an explicitly stated interpretive argument, it is easy to make claims based on implicit assumptions. For example, a high-school graduation test might be touted as a predictor of future performance in school and life, but be validated as a measure of a limited

domain of content knowledge. As Shepard (1993) pointed out, a test label like "developmental maturity" can "smuggle in whole theories without test users being aware of the choices they have made" (Shepard, 1993, p. 425).

2.5. Informal, Presumptive Arguments

In discussing validity arguments, Cronbach (1988) focused on the role of "reasonable argument based on incomplete information" (p. 5). Mathematical models can carry us part of the way and provide a comfortable aura of objectivity and certainty, but many of the core issues in the interpretations and uses of test results take us beyond the safe harbor of strictly deductive inference. As noted earlier, the evaluation of the interpretive argument for a placement test typically involves judgments about the appropriateness of the test tasks to be used, about the appropriateness of scoring criteria, about the relevance of the test tasks to the competencies emphasized in the courses, and about the value placed on completing a higher-level course relative to the risk of failure in a course that may be at too high a level. As a result, interpretive arguments tend to rely on *presumptive or informal reasoning* (Blair, 1995; Pinto, 2001; Toulmin, 1958, 2001; Walton, 1989) as well as mathematical or logical reasoning. Appraisals of presumptive, informal reasoning focus on coherence and on the plausibility of inferences and assumptions.

Formal reasoning is considered *valid* (in the logical sense) if the conclusions can be generated automatically from the premises using rules of inference. For example, the implications of a mathematical model can be derived more or less automatically once the model is accepted and values are assigned to parameters and initial conditions. Such formal reasoning plays an essential but limited role in interpretive arguments. Given that we accept their assumptions, various statistical, psychometric, and cognitive models can be used to draw conclusions from test performances. However, the application of a model to test scores always involves assumptions that are subject to doubt. Real-world arguments (including those based on scientific theories) have to attend to the question of whether the model fits the data and is appropriate given the proposed use. For example, any evaluation of the use of a test score to predict performance in a course involves the choice (implicit or explicit) of a criterion of course performance.

Interpretive arguments are informal and presumptive. They include inferences that do not follow formal rules of inference, and as a result, they exhibit three important characteristics of informal, presumptive arguments. First, their substantive assumptions and conclusions are subject to empirical challenge, and they are, to a large extent, evaluated in terms of how well they stand up to such challenges. If a presumptive argument has survived serious challenges, it is reasonable to have some confidence in its conclusions. The extent to which proposed interpretations and uses of test scores can stand up to serious criticism is a central issue in validation (Cronbach, 1971; Messick, 1989; Moss, 1998a, 1998b).

Second, the claims made by interpretive arguments and other presumptive arguments are always somewhat tentative and often include explicit indications of their uncertainty. Claims based on even highly confirmed scientific theories are subject to some uncertainty because of ambiguity in initial conditions and because all theories are approximations on some level. Many of the inferences in interpretive arguments (e.g., extrapolations from placement test scores to expected performance in a course) are subject to potentially large errors, and the uncertainties in their conclusions are often indicated by correlations, standard errors, confidence intervals, information functions, or Bayesian posterior distributions.

Third, like all presumptive arguments, interpretive arguments are *defeasible* in the sense that even when they are accepted in general, they can be overturned in a particular case (Pinto, 2001; Toulmin, 1958). That is to say, "there are possible facts ... that would cancel the presumptive support if they should come to light" (Pinto, 2001, p. 105). Even if the generic form of an interpretive argument is strongly supported, the conclusions may be overturned if they are contradicted by reliable evidence (e.g., if a student placed in an advanced course cannot keep up with the pace of the course).

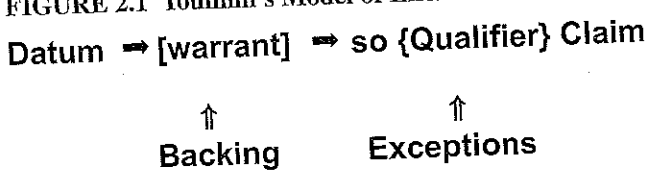
Informal, presumptive arguments can establish *presumptions* in favor of conclusions or decisions, but the inferences are not mechanical or certain. Like scientific theories, even well established informal arguments can be challenged in particular cases. However, by establishing a presumption in favor of certain claims, presumptive arguments shift the "burden of proof" onto those who would challenge its claims.

2.5.1. Toulmin's Model of Inference

Toulmin (1958) suggested a general framework and terminology for analyzing arguments, which has been widely used in a variety of contexts including program evaluation (Fournier, 1995) and measurement theory (Mislevy, 1996; Mislevy, Steinberg, & Almond, 2003). According to Toulmin (1958), the assertion of a claim (e.g., a conclusion or decision in an interpretive argument) carries with it a duty to support the claim and defend it if challenged, that is, to "make it good and show that it was justifiable" (Toulmin, 1958, p. 97). For an interpretive argument applied to test scores, this general principle of argumentation is reinforced by the *Standards for Educational and Psychological Testing* (AERA et al., 1999), which require that the claims based on test scores be supported by appropriate evidence.

Toulmin's model, which is represented in Figure 2.1, applies to the individual inferences within an argument. Each inference starts from a *datum* (D) and makes a *claim* (C). The initial inference in a quantitative interpretive argument is likely to be from a record of performance on some tasks (the datum) to a score (the claim). Subsequent inferences may involve generalizations or extrapolations of these scores, explanations of the scores, and decisions. Each interpretive argument is likely to involve a number of inferences, with the conclusions or claims (C) of earlier inferences serving as starting points, or data (D) for later inferences. The interpretive argument for placement tests outlined in Table 2.1 contains four inferences, with the

FIGURE 2.1 Toulmin's Model of Inference



Note: Backing provides the evidence for a warrant. Exceptions indicate conditions under which an otherwise sound inference may fail.

conclusions of each of the first three inferences serving as the datum for the following inference.

Each inference is made using a *warrant* (W), which is a rule for going from D to C. Toulmin suggests that this datum-warrant-claim structure is quite general, but that the warrants are specific to different kinds of inferences. The warrants used in interpreting test results are quite varied. The warrant for the scoring inference is the scoring rule. The warrant for generalization from the observed score to the universe score is a statistical generalization from a sample mean to the expected value over the universe from which the sample is drawn. The warrant for extrapolating to expected performance in a course may be a regression equation. The decision rule for the placement system provides a warrant for assignments to particular courses.

Warrants are not generally self-evident, and therefore, they have to be justified. The evidence supporting the warrant is the *backing* (B) for the warrant. The backing for scoring rules generally consists of expert judgment. The backing for generalization would include evidence that the sample is large enough and representative of the universe. Empirical studies of the relationship between test scores and criterion scores provide the backing for a regression equation. The justification for a decision rule typically relies on an analysis of the likely consequences (positive and negative) of using the rule.

Toulmin (1958) includes two additional components in his analysis of inferences, both of which are relevant to interpretive arguments and their validation. First, he allows for the inclusion of a *qualifier* that indicates the strength of the claim. Many of the inferences employed in the interpretations of test scores have explicit qualifiers. Standard errors and their associated confidence intervals indicate the uncertainty in generalizing from an observed score to a universe score. Inferences from a test score to a criterion score are usually accompanied by standard errors of estimation and/or correlation coefficients. Score-based decisions also tend to have qualifiers indicating their strength. They can have various levels of force ranging from a suggestion (e.g., in a career guidance system), to a strong suggestion (e.g., in a course-placement system), to a firm decision (e.g., in a college admission or licensure decision).

Second, Toulmin's model of argument allows for the specification of exceptions, or *conditions of rebuttal*, indicating conditions under which the warrant (which is defeasible) would not apply. In the context of validity, the exceptions would involve certain cases or categories of cases in which the interpretive argument is not justified.

Some possible exceptions may be explicitly included in the description of the testing program and its associated interpretive argument. The testing procedures and the proposed interpretations may be developed for specific populations, defined by age, educational background, language proficiency, etc. The use of the test with individuals who are not in the specified population might not be consistent with the assumptions built into the interpretive argument. In addition, certain possible contingencies that could interfere with the proposed interpretation may be explicitly identified. For example, a testing program that employs an IRT model as an integral part of its interpretive argument might routinely apply tests of person fit to all examinee's responses (Yen and Fitzpatrick, this volume). An inference from an observed score to a latent variable defined by the IRT model would be suspect for any examinee whose responses do not fit the IRT model.

An important class of exceptions to the applicability of standardized testing procedures and therefore standard interpretive arguments involves examinees with disabilities (Hansen, Mislevy, & Steinberg, 2003; Kane, 1992). Most high-stakes testing programs routinely provide accommodations for individuals with certain disabilities, thus replacing the standard interpretive argument with a modified interpretive argument that takes the accommodations into account. The goal is to reach the same kind of conclusions for all students, and the testing accommodations are designed to achieve this goal (Sireci, Scarpati, & Li, 2005).

In addition to any explicitly stated exceptions, every interpretive argument includes a general assumption to the effect that nothing has interfered with the proposed interpretation. Within the philosophy of science, this kind of general caution is referred to as the *ceteris paribus* (or "all else being equal") assumption. There are many unusual circumstances that could make the "else" unequal enough to interfere with the proposed interpretation and thus to generate an exception. For example, an inference from scores on a reading test to conclusions about a student's reading level may hold ordinarily (i.e., the relevant warrant may have strong empirical and theoretical backing), but not apply to a very farsighted student with broken glasses. Unusual circumstances like this would not ordinarily be anticipated in the interpretive argument, but they would be covered by the ceteris paribus assumption.

Toulmin (1958) treated his model as a dialogue between an advocate for a claim and a challenger. The advocate makes a claim based on an inference. The challenger can question the warrant for the inference or the appropriateness of applying the warrant in a particular case. If the warrant itself is challenged, its backing can be brought forward. The nature of this backing will depend on the nature of the warrant and the nature of the objection to it. For example, if the warrant is a regression equation linking test scores to expected performance in a course, its backing could consist of an empirical study relating test scores to course grades. The challenger can question the quality of the study, and the proponent can respond by defending the study or by questioning specific claims made by the challenger (and thereby shifting the burden of proof to the challenger). A challenger who accepts the warrant can still claim an exception in a

particular case (based on special circumstances), and the proponent can agree or argue the point.

The validity argument is to provide an overall evaluation of the evidence for and against the proposed interpretation. Assuming that an interpretive argument lays out the inferences involved in getting from the test scores to the conclusions to be drawn and the decisions to be made, the validity argument would provide a critical appraisal of the warrants and backing for these inferences.

When an inference is drawn (e.g., from an observed performance to a score), the warrant (e.g., the scoring rule) and its backing (judgments by panels of content experts who developed the scoring rule) may not be explicitly mentioned, especially if the inference is fairly routine in a particular context and the audience is friendly. Although they are generally not included in routine score reports, the warrants are an integral part of the interpretive argument and could presumably be supplied if needed. Similarly, the possible exceptions to the interpretive argument would not ordinarily be included in a score report, unless the exception applies to the report and makes a substantive difference in the interpretation or in the confidence with which the interpretation can be stated. The backing for various warrants, qualifiers, and possible exceptions would be included in a technical report or other supporting documentation.

2.6. Criteria for Evaluating Interpretive Arguments

Although they cannot be proven, interpretive arguments can be rigorously evaluated against general criteria for sound presumptive arguments.

Clarity of the argument. The interpretive argument should be clearly stated as a framework for validation. The inferences to be used in getting from the observed performance to the proposed conclusions and decisions, as well as the warrants and backing supporting these inferences, should be specified in enough detail to make the rationale for the proposed claims apparent. Implicit assumptions can be particularly harmful because they may be left unexamined.

Coherence of the argument. The argument is expected to be coherent in the sense that the network of inferences leading from the observed performances to conclusions and decisions makes sense assuming that the individual inferences are plausible. It is also expected to be complete in the sense that no essential inferences or assumptions are left out (AERA et al., 1999; Crooks, Kane, & Cohen, 1996).

Plausibility of inferences and assumptions. The assumptions included in the interpretive argument should be plausible. Some assumptions may be taken for granted, some can be supported by careful documentation and analysis of procedures (e.g., sampling assumptions), and some require empirical evidence to be considered plausible. For highly questionable inferences or assumptions, it is appropriate to consider several parallel lines of evidence. The plausibility of an assumption is judged in terms of all of the evidence for and against it.

Note that procedural evidence cannot go very far in establishing the validity of most interpretations, but it can be decisive in refuting an interpretive argument. If the procedures have not been followed correctly or if the procedures themselves are clearly inadequate, the interpretive argument would be effectively overturned.

If it were necessary to support every inference and assumption with empirical studies conducted after the assessment procedures are developed, validation would be essentially interminable, because most interpretations involve a number of inferences each of which relies on multiple assumptions, and the evaluation of these assumptions will rely on other assumptions. Fortunately, many inferences and assumptions are sufficiently plausible a priori to be accepted without evidence unless there is some reason to doubt them in a particular case.

Like scientific theories, interpretive arguments can be challenged in various ways, but one of the most effective ways to challenge a scientific theory or an interpretive argument is to propose an alternative theory or argument that is more plausible. The evaluation of plausible competing interpretations is therefore an important component in the evaluation of any proposed interpretive argument. The validity argument can make a positive case for the proposed interpretations and uses of scores by providing adequate backing for the interpretive argument and by ruling out plausible alternative interpretations.

2.7. Specifying the Proposed Interpretation

Score interpretations are constructs in the sense that they are constructed by some person or persons, and a test score may have several legitimate interpretations and may be used to make different kinds of decisions, each with its own interpretive argument. The evidence required for validation depends on the proposed interpretation, and it is entirely possible for one or more of these interpretations to be valid, while other interpretations are considered invalid. For example, it is possible that the test scores provide good indications of skill in performing some task but do not support any broader interpretation. The test may provide a good indication of achievement in an area but not be useful for placement testing. The test, the population of examinees, and the context may all remain the same, and yet, validity will vary from one interpretation to another:

> The proper goals in reporting construct validation are to make clear (a) what interpretation is proposed, (b) how adequately the writer believes this interpretation is substantiated, and (c) what evidence and reasoning lead him to this belief. (Cronbach & Meehl, 1955, p. 297)

If the interpretations and uses are not clearly specified, they cannot be adequately evaluated (Cronbach, 1989; Linn, 1998; Ryan, 2002).

In a sense, each interpretive argument is unique, and therefore, the associated validity argument is also unique. A validity argument evaluates a particular interpretive argument in a particular context. Although placement systems have many features in common, each placement program is unique in the instructional options available, the students to be placed, and the consequences of different placements. The appropriateness of a placement program depends on how well it works in a particular context and

its appropriateness for any particular student is contingent on that student's satisfying the assumptions built into the system.

The next four sections of this chapter present analyses of four broad categories of interpretive arguments: trait interpretations, theory-based interpretations, qualitative interpretations, and decision procedures. The discussion of specific validation methods is embedded within the analyses of these four categories of interpretive arguments in order to emphasize that the issues to be addressed in a validity argument are determined by the interpretive argument being evaluated.

3. TRAITS

A *trait* is a disposition to behave or perform in some way in response to some kinds of stimuli or tasks, under some range of circumstances. Much of the meaning of the trait is given by the domain of observations over which the disposition is defined, but trait interpretations also assume, at least implicitly, that some underlying or latent attribute accounts for the observed regularities in performance (Loevinger, 1957). For example, some individuals tend to get angry fairly easily, while others can put up with a lot of frustration without losing their tempers. Based on such consistent differences, we postulate an aggressiveness trait. Some students know a lot about chemistry, can answer various questions about this discipline, and can conduct experiments competently, while others cannot generally perform these tasks. We describe these differences in terms of level of proficiency in chemistry.

Messick defined a trait as: "a relatively stable characteristic of a person ... which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances" (Messick, 1989, p. 15). Trait language tends to be implicitly causal, but no specific mechanisms describe how the trait influences performance or behavior.

Some authors have focused on the domain of observations and see the underlying attribute as shorthand for patterns in the data:

> Much of psychological theory is based on trait orientation, but nowhere is there any necessary implication that traits exist in any physical or physiological sense. It is sufficient that a person behave as if he were in possession of a certain amount of each of a number of relevant traits and that he behave as if these amounts substantially determined his behavior. (Lord & Novick, 1968, p. 358)

Others focus on the underlying attribute and use the domain of observations as a vehicle for clarifying the meaning of the latent attribute:

> The evidence is very strong that there is a "general" factor of intelligence that is involved in a great variety of cognitive tasks.... Whenever a task requires noticing similarities and differences among elements, inferring correspondences, rules and generalities, following a line of reasoning, and predicting consequences, that task is likely to involve the general factor of intelligence—particularly as the elements of the task become more numerous and complex. (Carroll, 1986, p. 53)

Lord and Novick emphasize the observations and their relationships. Carroll emphasizes the underlying attribute. In both cases, the interpretation involves both a domain of observations and an underlying attribute.

3.1. Potential Circularity of Trait Interpretations

Explanations in terms of traits can be circular (Meehl, 1950). The trait is defined as a tendency to behave in particular ways under some circumstances and then the observed regularity is explained by the existence of the trait. Some people tend to get angry easily because they are aggressive and we know that they are aggressive because they tend to get angry easily. Some students can answer questions about chemistry because they are proficient in chemistry, and we know that they are proficient because they can answer the questions. As Cureton put it:

> We must not say that his high score is due to his high ability, but if anything the reverse. We say he has high ability because his performance has yielded a high criterion score. His "ability" is simply a summary statement concerning his actions. (Cureton, 1951, p. 641)

The use of trait language does not necessarily buy us much, and it can be misleading. It can suggest that we have found an explanation for an observed regularity, when we have merely labeled it.

In one sense, there is nothing objectionable about this kind of circularity. The observed regularities in performance are potentially important. We can use this regularity to make predictions about future performance, and it would be reasonable to cite the observed regularity as the basis for the predictions. If talk about traits is nothing more than shorthand for this kind of inference from past experience to expectations about the future, it is a legitimate form of shorthand (Meehl, 1950).

In most cases, however, we do have some sense of component processes involved in the trait, and therefore, the interpretation of the trait does not depend entirely on the performance domain. If nothing else, we know how we feel when we get angry, and we know how we would attempt to answer questions about chemistry. Trait interpretations involve general assumptions about the nature of the trait, and trait language reflects at least a rudimentary effort at theorizing. As a hypothesis about the person, a trait interpretation can be quite useful in interpreting experience.

3.2. Interpretive Arguments for Traits

The conception of a trait, as it is used here, encompasses a wide range of attributes, all of which involve a disposition to behave or perform in some way.

3.2.1. Target Domains

A trait is associated with a *target domain* of possible observations, and a person's expected score over the target domain is the person's *target score*. Some target domains are defined very broadly (e.g., Carroll's definition of intelligence); others are more circumscribed but still broad

(e.g., proficiency in algebra), and some are quite narrow (e.g., skill on a specific task). Target domains also vary in the extent to which they allow for varied conditions of observations. Some traits purport to describe how individuals will react in certain situations (e.g., test anxiety); others place few if any restrictions on context (e.g., literacy).

The decision to include certain observations in the target domain, and not other observations, generally relies on experience or prior assumptions about the processes involved in the observations. Certain tasks (e.g., arithmetic items) are included in a target domain because they are thought to require the same or at least overlapping skills or component performances. For traits as defined here, no explicit explanatory model is put forward, but general assumptions about the underlying attribute associated with the trait play a significant role in interpreting the trait scores and in defining the boundary of the target domain.

The target domains of most interest in education are not restricted to test items or test-like tasks, although they may include this kind of formal performance as a subset. A person's level of literacy in a language depends on his or her ability to perform a variety of tasks in a variety of contexts, ranging from the casual reading of a magazine to the careful study of a textbook or technical manual. These performances can occur in a variety of locations and social situations. The fact that it might be difficult to include some of these tasks in an assessment does not indicate that they are not part of the target domain for the trait. The match between the target domain and the measure of a trait is a central issue in developing a trait measure and in validating a trait interpretation.

The target domain may be somewhat fuzzy, in the sense that there are marginal cases in which it is not clear whether an observation should be included in the target domain. Whether a particular piece of technical prose should be included in the definition of literacy may be debatable, but in most cases it is not difficult to decide what's in and what's out. The point is not to define a tidy domain or one that is easy to assess, but to identify the range of observations associated with the attribute of interest.

3.2.2. Measurement Procedures

In some cases, it may be feasible to draw a random or representative sample of observations from the target domain and generalize a person's observed score on this sample to the person's expected score over the target domain. In most cases, however, the measurement procedure is standardized by restricting some conditions of observation. As a result, the range of observations included in the assessment is often much narrower than the range in the target domain. Restrictions may be imposed to promote fairness and replicability (i.e., reliability), for practical reasons (e.g., time limits), or for safety reasons. Consequently, the observations included in measures of traits are typically drawn from a subset of the target domain, often a very small subset (Fitzpatrick & Morrison, 1971; Kane, 1982).

The domain from which the observations are actually sampled by a measurement procedure is referred to as the *universe of generalization* for the measurement procedure, and a person's expected score over the universe of generalization is the person's *universe score* (Brennan, 2001a, 2001b; Cronbach et al., 1972; Shavelson & Webb, 1991).[1]

For standardized tests, the universe of generalization is a restricted subset of the target domain. Some aspects of the measurements may be uniquely specified. The format of the test and the instructions given to test-takers may be specified in detail. Some conditions of observation are specified by general guidelines; different forms of a test may include different sets of items, all following the same test plan. Some conditions of observations may become relevant only if they are extreme; for example, the environment in which observations are made may not be specified in any detail but can become an issue if conditions are extreme (very hot or cold, crowded, noisy, etc.).

While the target domain for adult literacy would include a very wide range of written material (e.g., novels, instructional manuals, magazines, memos, signs), responses (answering specific questions, giving an oral or written summary, taking some action based on the manual or sign), and contexts (e.g., at home, in a library, at work, or on the road), the universe of generalization for a measure of literacy may be limited to responses to objective questions following short passages while sitting at a desk or computer terminal. In most contexts, the reader can start and stop at will; in the testing context, the reader is told when to begin and when to stop. The performances involved in answering questions based on short passages under rigid time limits are legitimate examples of literacy but they constitute a very narrow slice of the target domain for literacy.

Standardization tends to be most extreme in objective tests, but it also occurs in performance testing (Fitzpatrick & Morrison, 1971; Messick, 1994). Students who are asked to conduct experiments for a science assessment are likely to be presented with a set of instructions, a setup including the necessary equipment and supplies, and a list of questions to be answered. The equipment presumably works or is quickly replaced. The work is to be completed under some time limits and without any unauthorized assistance. Actual scientific experiments are messier and longer term than most assessments can afford to be.

As a result of standardization, the samples of tasks included in measurements are not random or representative samples from the target domain, and it is not legitimate to simply generalize from the observed score to the target score. It is certainly not obvious, a priori, that performance on a passage-based objective test of literacy can be extended to the target domain for literacy, even if the observed scores are consistent over replications of the measurement procedure (and therefore generalizable over the universe of generalization).

The universe of generalization is usually a subset of the target domain. The interpretation of observed performance in terms of the target score requires a chain of reasoning from the test results to an observed score, from the observed score to the universe score, and from the universe score to the target score.

3.2.3. Trait Implications

Most trait interpretations carry implications that go beyond the target domain and that need to be addressed in

validation. A description "pulls behind it a whole train of implications" (Cronbach, 1971, p. 448), which may include expected relationships to other variables, the expected impact of interventions on the trait, and the extent to which differences are expected between groups.

The target domains for traits can provide a clearer and more definite meaning for existing concepts, like literacy, achievement in some academic subject, proficiency in some activity, or a disposition to behave in some way (e.g., personality traits). However, many trait labels were in use long before anyone decided to measure them, and they have implications that go beyond the specification of a target domain of observations (Bruner, 1990).

Assumptions about trait processes often imply relationships among traits. Two traits that involve similar or overlapping sets of processes are expected to be positively correlated. Traits that seem to involve very different processes would not ordinarily be expected to be highly correlated.

It may be assumed that the trait is or is not likely to change substantially over time. Some traits (e.g., general mental ability) are expected to remain quite stable at least for adults. State variables (e.g., moods) can change abruptly under certain circumstances. Proficiencies in specific areas (e.g., speaking French) can be expected to improve gradually as a result of effective instruction or practice on the activity.

For some traits, certain groups of people may be expected to get similar scores. For other traits, some groups may be expected to score higher than other groups. Experienced workers are expected to have a higher level of job proficiency than trainees. Individuals with certain diagnoses may be expected to have higher values of certain personality traits than individuals with no diagnosis or with a different diagnosis.

Trait measures are often developed for a particular application (e.g., to make placement decisions) and the label may suggest that the trait measure is appropriate for the application (Shepard, 1993). As Cook and Campbell (1979) pointed out, test developers "like to give generalized abstract names to variables" (p. 38), and as a result, trait labels may make implicit claims that the trait can be interpreted more broadly than would be suggested by the test development process.

Trait labels and descriptions typically involve values, as well as assumptions about the traits. Suppose, for example, that a target domain consisting of simple arithmetic tasks was used as the basis for a testing program. The label "test of arithmetic operations" has essentially no excess meaning. The label "test of basic skills in arithmetic" implies that essentially all students should be able to do well on the test. The different implications in the two labels deserve attention in evaluating the proposed interpretations and uses of the test.

A recurring theme in this chapter is the importance of stating the proposed interpretations and uses of test scores explicitly and evaluating all of the inferences and assumptions included in the interpretations and uses. Trait interpretations generally involve both expectations about performance over a target domain and general assumptions about an underlying attribute that accounts for observed regularities in the observations.

3.2.4. Traits as Unidimensional Attributes

Traits have their origins in attempts to account for differences among individuals in their performance in certain kinds of situations or on certain kinds of tasks. Individuals who are high on the trait are expected to do well on the tasks or to perform in a certain way in the situations. Those who are low on the trait are not expected to do well on the tasks or to perform in a different way in the situations. The observed score may vary from one observation to another and from one context to another, but the rank-ordering of individuals is expected to be similar across subsets of observations. In this sense, the trait is conceived of as a unidimensional attribute.

In practice, the unidimensionality assumption is mainly a statistical assumption rather than a substantive assumption about process (Lord & Novick, 1968). In many cases where trait terminology is used, it is known that the attribute being measured involves a combination of more basic components, but the attribute functions, at least approximately, as if it were a single, unidimensional trait in that "it operates in the same way as a determiner of success on all the items of the test" (Henrysson, 1971, p. 146).

3.2.5. Observable Attributes and Operational Definitions

As noted above, most traits have a dual interpretation, with part of their meaning given by the target domain and part of their meaning associated with an underlying attribute of the person. In some cases (e.g., in evaluating scientific theories), it is desirable to employ attributes that are largely devoid of excess meaning. Attributes that are explicitly defined in terms of a target domain of possible observations with few if any implications that go beyond the target domain are referred to as *observable attributes*. Observable attributes simply describe how well people perform some kind of task or how they respond to some kind of stimulus.

Operational definitions were developed to eliminate excess meaning altogether (Bridgeman, 1927). For operationally defined attributes, there are to be no implications that go beyond the universe of generalization defining the measurement procedure. Operational definitions are useful in some contexts, especially in empirical checks on theories. However, they have the potential to be misleading, particularly if the operational definition specifies a relatively narrow, highly standardized domain, but the label assigned to the score suggests a much broader domain or trait. For example, to define intelligence operationally as the score on a specific test but interpret it as a measure of overall cognitive functioning is to invite misinterpretation.

3.2.6. Quantitative Reasoning—An Example

Dwyer, Gallagher, Levin, and Morley (2003) define quantitative reasoning broadly, "as the ability to analyze quantitative information" (Dwyer, et al., 2003, p. 13), and specify that it involves the solving of quantitative problems that are new to the student. If the questions focus on reasoning that has been explicitly taught to some students, they are not included in the target domain for quantitative reasoning.

Dwyer et al. (2003) suggest that quantitative reasoning includes six capabilities: understanding quantitative information in various formats; interpreting quantitative information and drawing inferences from it; solving quantitative problems; estimating answers and checking them for reasonableness; communicating quantitative information; and recognizing the limitations of quantitative methods. Quantitative reasoning is defined in terms of performance on tasks that require these capabilities. The capabilities defining the domain are specified in more detail than is usually the case, but no explicit theory of performance is offered.

For Dwyer et al. (2003), content knowledge is considered an ancillary attribute, which is not part of the definition of quantitative reasoning and therefore should not have much impact on scores:

> the mathematical content in an assessment of quantitative reasoning should include only that which all test takers can be assumed to possess. (Dwyer et al., 2003, p. 17)

The interpretation assumes that the test tasks do not require specific mathematical or substantive knowledge that is unfamiliar to the students, and a finding that some students lack such knowledge could generate an exception to the proposed interpretation in terms of quantitative reasoning.

3.2.7. Overview of Interpretive Arguments for Traits

The interpretive argument implicit in trait interpretations is outlined in Figure 2.2. The interpretation of the trait and its relationship to its measurement procedure are represented in the middle and left side of the figure, and the interpretive argument is represented on the right side of the figure. The interpretive argument is stated in more detail in Table 2.2.

The target domain, representing the full range of possible observations associated with the trait, is in the center of the figure. The underlying attribute or trait is represented by the oval at the top of the figure. It is recognized that the observations in the target domain are also influenced by other traits, by the context in which the observations occur, and by the conditions of observation in potentially complex ways. These factors are represented by the ovals to the left

FIGURE 2.2 Measurement Procedure and Interpretive Argument for Trait Interpretations

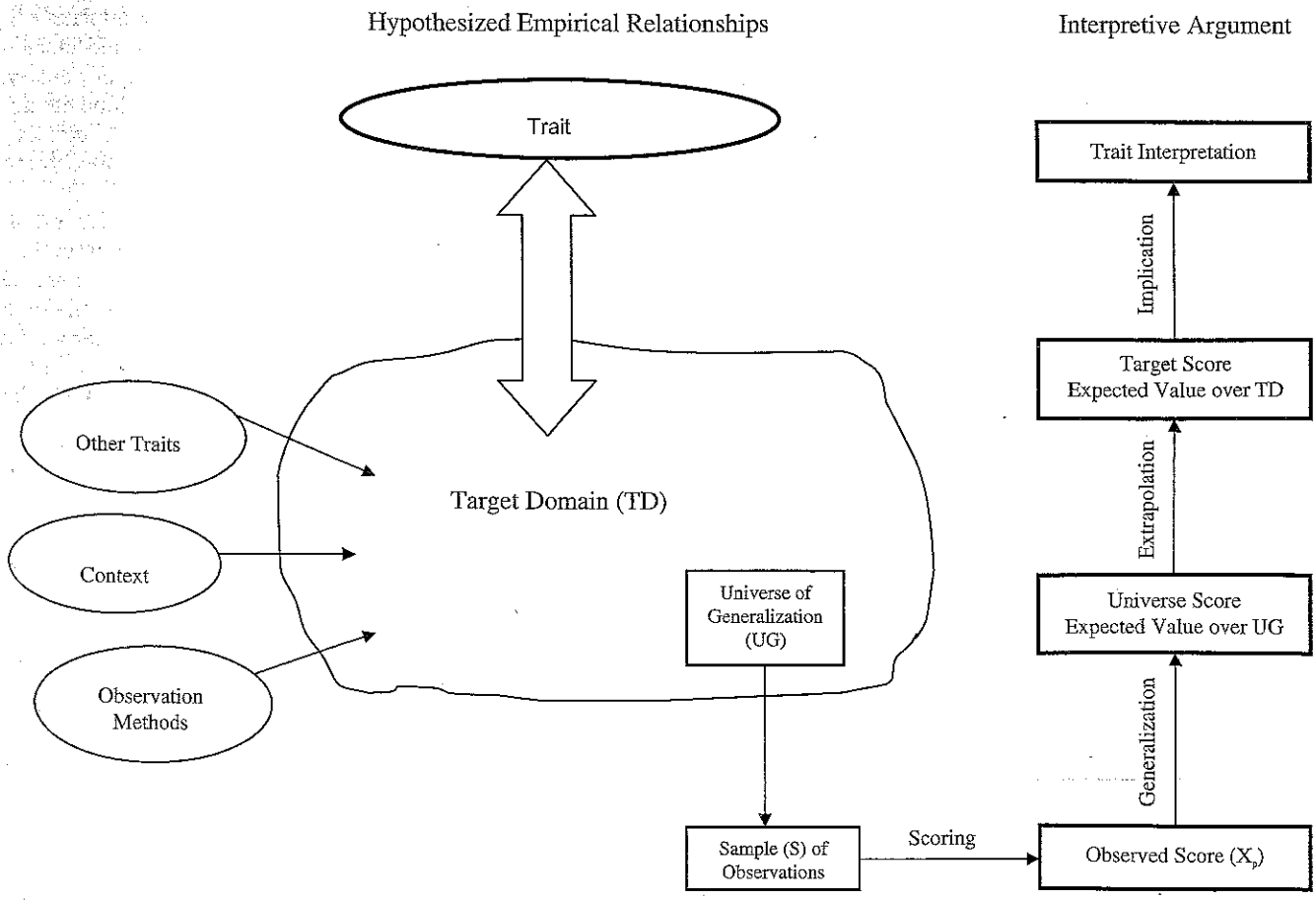


TABLE 2.2 Interpretive Argument for a Trait Interpretation

I1: ***Scoring***: from observed performance to the observed score
- A1.1 The scoring rule is appropriate.
- A1.2 The scoring rule is applied as specified.
- A1.3 The scoring is free of bias.
- A1.4 The data fit any scaling model employed in scoring.

I2: ***Generalization***: from observed score to universe score
- A2.1 The sample of observations is representative of the universe of generalization.
- A2.2 The sample of observations is large enough to control random error.

I3: ***Extrapolation***: from universe score to target score
- A3.1 The universe score is related to the target score
- A3.2 There are no systematic errors that are likely to undermine the extrapolation.

I4: ***Implication***: from target score to verbal description
- A4.1 The implications associated with the trait are appropriate.
- A4.2 The properties of the observed scores support the implications associated with the trait label.

of the target domain. The boundary of the target domain is typically not specified with great precision. As indicated in Figure 2.2, the universe of generalization for the measure of the trait is often a small subset of the target domain and tends to be defined more precisely than the target domain.

The interpretive argument on the right side of Figure 2.2 employs four major inferences in interpreting test results in terms of a trait. The observed performance is scored yielding an observed score (a raw score or scaled score of some kind), the observed score is generalized to the universe score, and the universe score is extrapolated to the target score. Finally, the implications associated with the trait are attached to the target score.

The scoring inference assigns a score to each person's performance using a scoring rule, which provides the warrant for the scoring inference. The specific criteria included in the scoring rule will depend on many factors, including the kinds of tasks included in the assessment and its purposes. The scoring inference relies on assumptions that the scoring criteria are appropriate and are applied as intended, that the process is free of bias, and that any statistical models (scaling, equating) employed in scoring are appropriate.

Generalization extends the interpretation of the observed score from an evaluation of a sample of observations to the expected value over the universe of generalization. The value of the score is the same, but its interpretation is expanded from a claim about a specific set of observations to a claim about expected performance over the universe of generalization. The warrant for this inference is derived from statistical sampling theory and depends on assumptions about the representativeness of the sample of observations and about the adequacy of the sample size for controlling sampling error.

Extrapolation from the universe of generalization to the target domain extends the interpretation from the universe score to the target score. The extrapolation inference assumes that the universe score is related (rationally and/or empirically) to the target score and that the extrapolation is relatively free of systematic and random error. Again, the score does not change, but the interpretation of the score is extended from the universe of generalization to the target domain.

The implications involve extensions of the interpretation to include any claims or suggestions associated with the trait. The warrants for trait implications provide authorization for any inferences stated or implied in the trait description, the trait label, and the uses made of scores. Many trait labels have rich associations and are highly value laden, especially those that have been part of the language for centuries. In adopting a preexisting trait label, a test developer is implicitly adopting this excess meaning as part of the proposed interpretation or is taking on an obligation to counteract any unwarranted inferences based on the trait label.

In order for the interpretive argument outlined in Table 2.2 to be convincing, each of the separate inferences must be convincing. A failure of any one of the inferences undermines the interpretive argument as a whole, even if the evidence for the other inferences is compelling (Crooks et al., 1996).

3.3. Validity Arguments for Trait Interpretations

The validation of a proposed interpretation begins with an overall evaluation of its coherence and completeness. For a trait interpretation, this initial review would include an evaluation of the appropriateness of the target domain, given the trait label and description (or conversely, the appropriateness of the label and description given the domain) and an evaluation of the coherence and completeness of the interpretive argument.

The interpretive argument outlined in Figure 2.2 and in more detail in Table 2.2 is intended to provide an outline of how trait measures are actually interpreted. The initial review would examine how well the proposed interpretive argument represents the case at hand. If the interpretive argument does not provide a reasonable explication of the proposed interpretations or uses of the test results, it should be replaced by a more appropriate interpretive argument.

3.3.1. Scoring

Much of the evidence for the scoring inference is based on the judgment of panels of experts who develop and review the scoring criteria, on the care with which the scoring procedures are implemented, and on the procedures used to select and train scorers (Clauser, 2000). These issues are typically addressed during test development.

Several kinds of empirical evidence may also be relevant to the scoring inference. Empirical data can be used to check on the consistency (e.g., interrater reliability) and accuracy (e.g., quality control data) of scoring. A demanding check on the scoring criteria would have two or more panels develop scoring rules independently and would then evaluate the agreement between the scores generated by the two rules (Clauser, Harik, & Clyman, 2000). For many standardized testing programs, statistical models are used

to scale or equate scores, and the fit of these models to the data can be evaluated empirically.

The factors that can undermine the warrant for scoring of the observed performance are as numerous as the things that can go wrong in administering and scoring an assessment. The scoring rubrics may reflect inappropriate criteria or fail to include some important, relevant criteria. The selection or training of the scorers may be flawed. Quality control procedures (e.g., checks on the scorers' consistency and accuracy in applying the scoring rubrics) may be inadequate. In addition, exceptions to the scoring inference can arise from violations of the specified procedures.

Procedural evidence cannot establish the plausibility of an interpretive argument, but it can be decisive in refuting it. If the procedures have not been followed correctly (e.g., the wrong scoring key was used) or if the procedures themselves are clearly inadequate (e.g., inadequate training of raters), the interpretive argument would be effectively refuted.

3.3.2. Generalization

The measurement procedure is designed to yield representative samples from the universe of generalization, and if the sample of observations is representative of the universe and is large enough to control sampling error, sampling theory provides a warrant for generalization to the universe score (Loevinger, 1957). If a serious effort has been made to draw a representative sample from the universe of generalization, and there is no indication that this effort has failed, it would be reasonable to assume that the sample is representative (Kane, 1996). As Eisner (1991) pointed out, sampling assumptions are rarely satisfied, and therefore, "inferences are made to larger populations, not because of impeccable statistical logic, but because it makes good sense to do so" (Eisner, 1991, p. 203).

Even if we could draw perfectly representative samples from the universe of generalization, estimates of the universe score would still be subject to sampling errors, and generalization of an observed score to the universe of generalization assume that these sampling errors are not too large. The empirical evidence needed to support generalization over replications of a measurement procedure is collected in reliability studies (Feldt & Brennan, 1989; Haertel, this volume) or generalizability studies (Brennan, 2001b; Cronbach et al., 1972). Any facet that is allowed to vary in the universe of generalization (e.g., tasks, occasions, raters) and is sampled by the measurement procedure contributes to random error. To the extent that reliability or generalizability studies indicate that the sampling errors associated with replications of the measurement procedure are large, inferences from the observed score to the universe score are uncertain.

We have at least two options if the random errors associated with a facet are large. First, during the development phase, we can modify the measurement procedure (e.g., by increasing sample sizes for the facet) so that sampling variability is reduced. Second, we can modify the definition of the attribute so that it does not involve generalization over the facet in question. For example, if the scores are not generalizable over certain kinds of tasks, those tasks could be excluded from both the target domain and the universe of generalization. This has the effect of narrowing the definition of the attribute and thereby may limit its usefulness (Kane, 1982).

Generalizability or reliability studies provide estimates of standard errors of measurement and therefore put limits on the precision of estimates of the universe score (Brennan, 2001a, 2001b). Large standard errors and broad confidence intervals imply weak conclusions about the universe score.

The ceteris paribus assumption plays an important role in claims for generalizability (Kane, 2002a). In particular, if the observations did not occur under conditions consistent with the measurement procedure, they would not constitute a sample from the universe of generalization. For example, if the conditions of observation involve impediments to performance (e.g., faulty equipment) or inappropriate aids to performance, the observations would not be representative of the universe of generalization.

3.3.3. Extrapolation

The extrapolation inference can be evaluated using two kinds of evidence, analytic and empirical. The analytic evidence relies on conceptual analyses and on judgments about the relationship between the universe of generalization and target domain. Much of the analytic evidence tends to be generated during the development stage, as the interpretive argument and measurement procedure are developed. The empirical evidence examines relationships between observed scores and other scores associated with the target domain (e.g., other measures drawn from the target domain).

3.3.3.1. Analytic Evaluation of Extrapolation Inference

Extrapolation depends, at least in part, on the relationship between the universe of generalization and the target domain. It is most plausible if the universe of generalization covers a large part of the target domain or employs high-fidelity simulations covering much of the domain (Flockton & Crooks, 2002). If the universe of generalization constitutes a highly restricted subset of the target domain, the inference from the universe score to the target score is generally more questionable.

The argument for extrapolation may also be based on general notions about overlap in the processes employed in responding to the test tasks and other tasks in the target domain (Snow & Lohman, 1989). For example, if the goal is to make claims about students' skill in using mathematics in a variety of real-world situations, and the test consists of a sample of problems involving the mathematical skills required in the real-world situations, it could be reasonable to assume that students who are successful on the test would also be successful in the real situations and that students who are not successful on the test would not be successful in the real situations.

The processes involved in the test tasks and in other tasks in the target domain can be identified using "think-aloud" protocols. These data could be collected in one-on-one sessions

with researchers recording the candidate's self-description of how they approach each task (Bonner, 2005; Cronbach, 1971). If the processes used in responding to test tasks are similar to those used in responding to other tasks in the target domain, confidence in extrapolation is enhanced. To the extent that the performances on the test tasks are substantially different from those for other tasks in the target domain, extrapolation is suspect.

In practice, analytic evidence for extrapolation helps to rule out threats to the credibility of this inference. If a serious effort is made to identify factors that could produce substantial differences between the universe score and the target score, and no such factors are found, the extrapolation is likely to be accepted. If the impact of some factors on the plausibility of extrapolation is unclear, it may be necessary to check on their importance empirically.

The notion of "face validity" refers to the apparent relevance of test tasks to the proposed interpretation or use of scores, and efforts to build face validity into tests are intended to enhance acceptance of the test by those taking it and by various stakeholders (e.g., parents, employers). The appearance of relevance does not go far in supporting the appropriateness of a trait interpretation, but a serious lack of such relevance can lend credibility to certain challenges to the extrapolation inference (Messick, 1989). In particular, to the extent that students put less effort into their performance on a test than they would on the corresponding tasks in other settings, because the test seems irrelevant, the extrapolation inference would be weakened.

Exceptions to the extrapolation inference would generally involve cases in which performance on the test is likely to be different from performance in other parts of the target domain. Any disability that interferes with test performance but not with performance in other parts of the target domain would generate exceptions to the extrapolation inference. Similarly, a lack of some ancillary skill (e.g., a high level of reading comprehension) that is required on the test but is not generally required in the target domain could introduce bias into the estimate of the trait, and could therefore generate an exception. Note however, that a disability or skill deficit that interfered to the same degree with performance on the test and in the target domain would not count against the extrapolation inference.

3.3.3.2. Empirical Evaluation of the Extrapolation Inference

The warrant for extrapolating to the target score could be evaluated empirically by comparing observed scores to criterion scores based on an especially thorough (and representative) sample of performances from the target domain. Criterion-related validity evidence seeks to establish such a direct link between test scores and a demonstrably valid criterion measure (Cronbach, 1971; Messick, 1989).

It is often possible to obtain estimates of the target score that are better than those provided by a standardized test by sampling more broadly from the target domain. Assume, for example, that a test of proficiency in a foreign language consists of sets of objective items asking questions about short printed passages in the foreign language and taped conversations in the foreign language. The question of how well the test represents the target domain for language proficiency can be addressed, at least in part, by developing a criterion measure that samples the target domain more thoroughly than the test. The criterion assessment might include one-on-one interviews, group discussions, or tasks involving the use of printed material, all in the foreign language. Assuming that the criterion measure is only going to be given to a relatively small number of subjects in a validity study, it can be time consuming and expensive.

3.3.3.3. Validity Generalization

Validity generalization involves the application of results from existing criterion-related validity studies to new situations, new populations of examinees, and possibly, to a new test that is similar to those in the existing studies (Murphy, 2003). For example, if several studies of criterion-related validity using a test to predict some criterion in some type of setting have been found to yield consistent results, it can reasonably be concluded that similar results will be found for the test in a new situation, which is similar to those in the earlier studies. The assumption that the new situation is similar in program characteristics, setting, and population to those in the previous studies is based on judgments about the characteristics that could make a substantial difference in the results.

The inference to a new situation or test is strengthened to the extent that it is based on a number studies rather than one or two studies. If all of the studies in a certain kind of setting have yielded consistent results, it would be reasonable to extend the conclusions to a new setting of that kind.

3.3.3.4. Convergent Validity Evidence

Given the difficulties inherent in drawing representative samples from broadly defined target domains, it is often not feasible to develop a clearly valid measure of the target score for use as a criterion measure. An alternate approach is to develop measures of the target score involving different kinds of standardization and therefore different universes of generalization.

Convergent validity evidence (Campbell & Fiske, 1959) is evaluated in terms of the correlations between different measures of a trait. If these correlations are low, at least some of the measures are not adequately representing the target domain. If the correlations of a given measure with other measures drawn from different parts of the target domain are high, extrapolation to the target domain is supported. Convergent evidence may be based on correlations between different tests developed for this purpose or between existing tests with the same label.

It is also possible to use convergent evidence to evaluate specific threats to extrapolation. For example, two teams of qualified test developers could be assigned to develop tests for the same test specifications. A high correlation between the independently developed tests suggests that the systematic errors introduced by assessment teams are small. A low correlation suggests that the test-development team introduces substantial systematic error (Cronbach, 1971).

There are many potential sources of systematic error in most measurement procedures (e.g., testing instructions, time limits, locations, equipment and materials, order of the questions). Standardization of any aspect of the measurement procedure that is not also fixed in the target domain introduces a source of systematic error (Kane, 1982). The goal in designing measurement procedures is to standardize in ways that control random error effectively while introducing as little systematic error as possible.

In practice, it makes sense to check on those aspects of standardization that are most likely to introduce appreciable error. The choice of assessment method (e.g., objective, essay questions, hands-on performance), of scoring rule (Clauser et al., 2000), and of the conditions of observation (Brennan, 2001b) may introduce irrelevant variance and therefore generate systematic error. Decisions about whether to investigate a particular aspect of standardization depend on a number of factors, including the nature of the measurement procedure, the proposed interpretation, previous research on invariance over conditions of the facet, and the stakes associated with the test.

3.3.3.5. Challenges to Extrapolation

Challenges to extrapolation claim that the universe of generalization is sufficiently different from the target domain that extrapolation from the universe score to the target score is not legitimate (Haertel & Greeno, 2003). To the extent that the assessment involves task presentations and/or response formats that are different from most of the tasks in the target domain (e.g., written responses in an area where most responses are oral or involve performances), irrelevant method variance may be introduced (Messick, 1989, 1994). Inferences from highly standardized measures to a broadly defined target domain are more prone to systematic errors than inferences from broadly defined measures (e.g., involving multiple formats, broad content sampling) to the full target domain.

Therefore, it is desirable that the universe of generalization cover as much of the target domain as possible. But there is a downside to this strategy. Performance tasks tend to be time consuming, and therefore, performance assessments tend to include a relatively small number of performances. As a result, generalization from the small sample of high-fidelity performances to a broadly defined universe of generalization may be quite undependable. There is a clear tradeoff here. We can strengthen extrapolation at the expense of generalization by making the assessment tasks as representative of the target domain as possible, or we can strengthen generalization at the expense of extrapolation by employing larger numbers of highly standardized tasks (Kane, Crooks, & Cohen, 1999). The goal is to strike a compromise that supports both generalizability and extrapolation.

Greeno, Pearson, and Schoenfeld (1997) argue that testing is a very specialized activity, and that

> Success in that situation depends on abilities to participate in the practices of test taking, which differ fundamentally from other practices that students need to be learning .... We create circumstances in which we can make reliable observations, but in the larger framework, those observations tell us about something that is relatively trivial. (Greeno, Pearson, & Schoenfeld, 1997, p. 170)

Taken literally, this perspective seems to say that extrapolation from standardized test contexts to other contexts is impossible in principle. As a general methodological assumption, this seems overly restrictive. As a warning about the risks inherent in casually extrapolating from standardized testing contexts to work, school, and life, it is a useful caveat (Haertel & Greeno, 2003).

3.3.3.6. Overall Evaluation of the Extrapolation Inference

Generally, a thorough evaluation of the extrapolation inference will require both analytic evidence and empirical evidence. Analytic evidence can be developed during test development by making the test as representative of the target domain as possible. To the extent that standardization is needed to control random error, and therefore the universe of generalization is not representative of the target domain, it may still be possible to design it to tap the core component skills involved in the target domain. For example, in measuring quantitative reasoning as defined by Dwyer et al. (2003), the test tasks could be designed to elicit the six capabilities specified in their framework and to require relatively little specific mathematical or substantive knowledge. In addition, an effort could be made to identify and control the most serious sources of systematic error.

During the appraisal stage, various kinds of empirical evidence that might challenge the proposed interpretive argument would be examined. For example, it might be possible to conduct a criterion-related study by administering a more thorough assessment of the attribute of interest to some sample of students and comparing scores on the test to scores on this more thorough criterion measure.

Convergent evidence might be obtained by comparing the test scores to scores on other measures of the trait. These other measures are not necessarily any better than the test, but they are different from the test and provide an empirical check on extrapolations from test performances to other performances in the target domain.

If the development stage yields a good preliminary case for extrapolation, and this case survives serious empirical challenges, it would be reasonable to accept the extrapolation inference. Under these circumstances, the interpretation of observed scores in terms of expected performance over the target domain would constitute a reasonable presumption.

3.3.4. Trait Implications

The trait label and description and any proposed uses of the test scores generally carry implications that go beyond the definition of the target domain. These implications extend the interpretation beyond a simple inductive summary.

Initial confidence in these implications is likely to depend at least in part on the fit between the target domain and the assumptions implicit in the trait (Cook & Campbell, 1979). If the target domain involves observations that are closely associated with the conception of the trait, and the most likely sources of systematic error have been ruled out, it would be reasonable to draw conclusions based on the conception of the trait. Most analytic evidence for trait implications is generated during the development phase, as the trait interpretation is specified and the test is designed to reflect the trait in terms of content, task types, procedures, context, and scoring (Dwyer et al., 2003; Loevinger, 1957).

Specific implications of the trait interpretation can also be checked empirically. If the conception of the trait suggests that it should be related to some other variable in some way, confirmation of this implication would support the proposed interpretation and disconfirmation would be evidence against the interpretation.

If the trait is expected to remain stable over time, empirical results suggesting that this is so would support the interpretation, and contrary results would undermine it. On the other hand, if the trait is expected to vary as a result of some intervention, change in the expected direction would support the proposed interpretation and stability would count against the interpretation.

For some traits, different groups might be expected to have substantially different score distributions. Individuals with an existing clinical diagnosis might be expected to exhibit extreme scores on some personality traits. Students who read well in class would be expected to perform better on a literacy test than students who are not reading well. These analyses based on categorizations are essentially natural experiments. They take advantage of the fact that the conception of the trait, general as it is, suggests certain patterns in the data.

Score differences between samples of students with similar interests, instructional backgrounds, and overall levels of achievement, but from different ethnic groups, would raise serious questions about the interpretation of the scores. More generally, indications that trait measures depend on any factor that is inconsistent with the conception of the trait would count against the validity of the measure.

To the extent that the implications associated with the trait are supported by data, the plausibility of the trait interpretation is enhanced. To the extent that these implications are disconfirmed, the plausibility of the trait interpretation decreases.

3.4. Trait Underrepresentation and Irrelevant Variance

As discussed in Section 2, challenges to the plausibility of presumptive arguments play a major role in the evaluation of such arguments. The two main threats to the plausibility of trait interpretations fall under the headings of underrepresentation and irrelevant variance (Cook & Campbell, 1979; Messick, 1989).

3.4.1. Trait Underrepresentation

A trait measure is said to under-represent the trait if it fails to adequately represent the range of observations or processes associated with the trait. From the point of view of sampling theory, standardized measurement procedures always underrepresent the target domain, because they treat some conditions of observation that vary in the target domain as fixed (e.g., fixed response formats). The extrapolation inference is supposed to bridge the gap between the universe of generalization and the target domain by indicating that this underrepresentation does not have any substantial impact on scores. For validation, serious underrepresentation occurs when restrictions in the universe of generalization relative to the target domain interfere with the estimation of the target score.

The extent to which extrapolation remains plausible over time can depend, in part, on the stakes associated with the test. In a low-stakes context, performance in various subsets of the target domain may represent performance in the target domain as a whole fairly well. In a high-stakes context, standardization to a subset of the target domain may lead to instruction and test preparation activities aimed specifically at the test, thus making test performance less representative of performance in the target domain as a whole.

The risks posed by different kinds of underrepresentation can often be prioritized based on experience and general assumptions about the processes involved in the trait. The more serious threats can then be evaluated empirically by conducting studies in which the conditions standardized in the measurement procedure are allowed to vary. For example, the impact of response formats could be examined by varying the format across different tasks.

In general, it is desirable for trait measures to include broad sampling of observations, both in terms of the processes involved and in terms of the tasks, situations, and contexts included (Cook & Campbell, 1979; Loevinger, 1957; Messick, 1989). By including a broad sampling of the observations associated with the trait, confidence in extrapolations from scores to trait values is enhanced. The use of multiple measurement methods helps to control irrelevant variance by averaging out method effects. Cook and Campbell (1979) and Messick (1989) raise strong concerns about "mono-operational bias" (Messick, 1989, p. 35) resulting from the use of a single observing method to estimate a trait.

Underrepresentation of the trait can constitute a general threat to validity or it can represent a possible exception for some individuals or groups. For example, a test involving novel problems (e.g., maximizing the area enclosed by a certain length of fencing under some constraints) might provide a good measure of problem solving among eighth graders. However, for students who have been taught how to solve such problems, the test would provide a measure of skill in applying these algorithms rather than a measure of problem solving.

3.4.2. Irrelevant Variance and Systematic Error

Trait measures are said to include irrelevant variance to the extent that they are influenced by factors that are not

associated with the trait. Concerns about irrelevant variance have a long history under the label of systematic error. There are a number of general sources of irrelevant variance, including traits other than the trait of interest, characteristics of the measurement procedures (e.g., rater bias, item parameter drift, misspecification of a psychometric model), and conditions of observation that affect scores (Downing & Haladyna, 2004; Haladyna & Downing, 2004).

Messick (1989, p. 34) defines construct-irrelevant variance in terms of "excess reliable variance that is irrelevant to the interpreted construct." Therefore, random error is not considered part of construct-irrelevant variance, even though random error variance is construct irrelevant. It is useful to maintain a distinction between random and systematic error, because the steps taken in controlling these two kinds of error are quite different. Random error is generally controlled by standardizing the measurement procedure and by collecting larger samples of performance in the assessment. The systematic errors that produce irrelevant variance have to be addressed in a way that is tailored to particular sources of error.

Standardization tends to promote fairness and to control random error, but it also transforms some sources of random error into sources of systematic error (Kane, 1982). For example, setting time limits on tests eliminates one potential source of random error and promotes fairness across testing sites; however, if speed is not included in the trait definition and individuals vary in the speed with which they work, the imposition of tight time limits can introduce systematic error into the trait measures.

Limiting the test to one method of assessment for a broadly defined trait can lead to both underrepresentation of the trait and irrelevant method variance. These problems can be alleviated by including multiple task formats (Messick, 1989, p. 35). If the observations are sufficiently diverse, "the errors are uncorrelated and more or less cancel each other out" (Loevinger, 1957, p. 648).

Note that an effect might be considered a source of irrelevant variance in measuring a trait but not be considered irrelevant with respect to criterion prediction, "because the criterion *measure* might be contaminated in the same way" (Messick, 1989, p. 34) [emphasis in original]. In this vein, Loevinger (1957) suggested that the empirical keying of items to optimize correlation with some external criterion (psychiatric diagnoses, job success) can produce a test with good predictive efficiency but with no clear trait interpretation. In this case, as in all cases, it is important to be clear about the intended interpretations and uses of test results.

Besides constituting a general threat to validity, irrelevant variance can create an exception for some individuals or groups. For example, a test may take some ancillary competencies (e.g., reading ability, computational skill) for granted. If these skills are variable enough in the population being tested to substantially influence scores, they can be considered sources of irrelevant variance. If the level of proficiency in an ancillary skill (reading English) required by the test is low compared to the levels of skill in the population, this would not constitute a source of irrelevant variance for most students but could be a serious source of systematic error for some students (e.g., those with limited English proficiency).

3.5. Profiles of Traits

In many cases, the focus of the interpretation is not a single trait but a profile of related traits, each with its own target domain and interpretive argument.

3.5.1. Discriminant Evidence for Divergent Traits

The traits included in the profile are expected to be clearly distinct from one another, and therefore the scores on different traits are not expected to be strongly related to each other. A strong relationship between measures of two distinct traits suggests that the two measures may be influenced by some common source of variance.

Empirical evidence on the relationship between distinct traits has been termed *discriminant validity evidence* (Cronbach, 1971). As is generally the case in validation, the role of discriminant evidence is shaped by proposed interpretations. If two measures are expected to be strongly related (e.g., because they are measures of the same trait), a strong empirical relationship between them provides evidence for the proposed interpretation. If two traits are expected to be unrelated, a strong empirical relationship counts against the proposed interpretation. The role of particular kinds of evidence in a validity argument depends on the proposed interpretive argument.

3.5.2. Multitrait-Multimethod Matrices

Campbell and Fiske (1959) proposed that the various kinds of correlational evidence relevant to the validation of a set of trait measures be considered within the context of a multitrait-multimethod matrix, like that presented in Table 2.3. By combining several issues into one analysis, these multitrait-multimethod matrices provide a richer set of conclusions than could be derived from several separate analyses investigating specific concerns.

The multitrait-multimethod matrix in Table 2.3 is based on hypothetical results. It assumes that a sample of individuals have been assessed on three traits, A, B, and C, using each of two methods, 1 and 2. This hypothetical matrix is about as simple as such matrices can be, but it presents evidence relevant to a number of important questions about the validity of the trait measures.

TABLE 2.3 Synthetic Multitrait-Multimethod Matrix for Three Traits and Two Methods

	Method 1			Method 2		
	A1	B1	C1	A2	B2	C2
A1	(.90)					
B1	.40	(.90)				
C1	.60	.40	(.90)			
A2	**.60**	.40	.55	(.70)		
B2	.40	**.65**	.40	.30	(.70)	
C2	.55	.40	**.60**	.65	.30	(.70)

First, the six entries in parentheses along the diagonal represent the correlations across independent measures of a single trait using the same method. Each of these entries can be considered an estimate of the reliability of the scores for that trait-method combination. Since the interpretive arguments for trait measures involve generalization from the observed score to the universe score for the measurement procedure, such generalizability is necessary for the interpretation to be viable. In addition, because limitations in the reliabilities of variables put limits on the correlations among these variables, some understanding of the reliabilities is needed for a clear interpretation of the multitrait-multimethod matrix.

The reliabilities for the three traits are all quite high for method 1 and substantially lower for method 2. One might expect to find the pattern of reliabilities in Table 2.3, for example, if method 1 were an objective test, and method 2 were a performance assessment.

Second, convergence of scores across methods for each of the three traits can be evaluated by examining the monotrait-heteromethod correlations, which are in boldface in the lower left of the matrix. Each of these three entries is the correlation between measures of a single trait using the two different methods. Ideally, these correlations would be large and positive. As a practical matter they are expected to be, "significantly different from zero and sufficiently large to encourage further examination of validity" (Campbell & Fiske, 1959, p. 81). The monotrait-heteromethod correlations in Table 2.3 are all quite high (.60 to .65). The maximum possible value for these correlations given the reliabilities is roughly .80.

Third, the requirement that different traits be distinct or discriminable suggests that the correlations between measures of different traits should not be very high. However, there is no reason to expect a priori that different traits will be independent of each other, and therefore, it is not possible to set any absolute requirement for how low the heterotrait correlations should be. Honesty and diligence are arguably distinct traits, but they are likely to be correlated in many populations. Therefore a fairly high correlation between measures of these two traits employing either the same method or different methods would not necessarily be surprising or distressing.

As Messick (1989) has suggested, investigations of the overlap among traits can be most useful for closely related traits:

> Empirically distinguishing the construct of assertiveness from sociability provides discriminant evidence, to be sure, but it is not as pertinent as distinguishing assertiveness from aggressiveness. (Messick, 1989, p. 48)

As a result, discriminant evidence tends to be most problematic when it is potentially most helpful.

The multitrait-multimethod matrix is particularly effective in providing reasonable relative standards for the discriminability of different trait measures. A modest requirement suggests that the correlations between measures of the same trait using different methods should be higher than the correlations between measures of different traits using different methods. This seems like a fairly minimal requirement, but as Campbell and Fiske (1959) point out, it is not always met. This requirement is met in Table 2.3. All of the heterotrait-heteromethod correlations (the off-diagonal correlations in the lower-left rectangle) are lower than the monotrait-heteromethod coefficients along the diagonal of this rectangle.

A more stringent standard would require that correlations between measures of the same trait using different methods be higher than the correlations between measures of different traits using a common method. This requirement is violated in Table 2.3. In particular, the correlation between Traits A and C for method 1 (.60) is as high as the monotrait-heteromethod coefficients for these variables, suggesting that they are not discriminated very well. The situation is even worse for method 2, where the correlation between traits A and C on method 2 (.65) is higher than the monotrait-heteromethod coefficients for these two traits (.60). This pattern could result from the "halo effect" or some other source of method variance, or it could result from a failure to clearly distinguish between the two traits.

Finally, Campbell and Fiske (1959) also suggest that the general pattern of correlations among the traits be similar across the different methods. Assuming that the different measures of each trait reflect mainly the trait and to a smaller extent the impact of method effects or random error, it would be expected that the pattern of relationships among the traits would be independent of the method. In practice, all measurements contain random errors, and some methods of measurement tend to involve larger random errors than other methods. The pattern of relationships among the three traits is fairly consistent in Table 2.3. All three traits are positively correlated. Traits A and C appear to be closely related, and trait B is relatively distinct from the other two. Overall, it appears that traits A and C are not being clearly distinguished, especially by method 2.

In those cases where they can be deployed, multitrait-multimethod matrices provide a systematic framework for looking at a range of validity issues. The emphasis on multiple methods and multiple traits provides effective criteria for evaluating the discriminability of traits and for evaluating the impact of systematic method effects by comparing correlations involving different methods to correlations involving the same method.

Note, however, that convergence across methods supports validity only if the proposed interpretation is independent of method. In discussing Campbell and Fiske's (1959) willingness to take the correlation of self report and peer ratings as evidence of convergence, Cronbach (1989) made the point that

> their justification could only be that *their* construction viewed self and others as interchangeable perceivers, equally willing to report what they see. For me, self-concept and reputation are distinct constructs, so I would find a high correlation troublesome rather than assuring. (Cronbach, 1989, p. 156)

The relevance of any kind of evidence to the validity of a proposed interpretation depends on the interpretation being proposed.

3.6. Factor Analysis

Traditional factor analysis seeks to model (and thereby to account for) the correlations among a set of variables using a relatively small number of underlying factors (McDonald, 1985; Spearman, 1927). Rather than associate test scores with a single underlying trait, factor analyses assume that these scores depend on a number of factors. They can therefore provide indications of the processes (i.e., factors) involved in the observed performances.

A factor analysis assumes that the test scores depend on a set of factors some of which are common to more than one test, and it uses the observed pattern of correlations among the tests to reason backward to estimates of how much the scores on each test depend on each of the hypothetical factors (i.e., the factor loadings for each factor). If no a priori assumptions are made about the factor loadings, the analysis is referred to as an *exploratory factor analysis*.

Using this kind of reverse inference, exploratory factor analysis seeks to identify a set of hypothetical factors, which can account for the observed pattern of correlations among test scores. The factors and factor loadings derived from the analysis are initially mathematical abstractions. The substantive meaning of the factors is developed by examining the content, format, and measurement procedures for the tests that load on each factor. For example, if all of the tests that load on a factor involve a fair amount of computation, and tests with little or no computation have consistently low or zero loadings on the factor, the factor could be identified with computational skill. The interpretation depends on a combination of formal mathematical modeling and subjective judgments that tie the model to observable phenomena.

The factors can be considered inductive traits, although the induction is not as straightforward as it is when a single trait is associated with a set of positively correlated task performances. The interpretation of the factors is developed inductively by observing the tests that load heavily on the factor, those with modest loadings, and those with no loading or negative loadings. As is true of other inductive traits, interpretations of test scores in terms of factors do not involve any detailed process model describing how the factor operates in producing the observed performances.

In *confirmatory factor analysis*, the general form of the factor structures for different attributes is specified in advance, and the data are used to evaluate the proposed model and to estimate the unconstrained loadings. For example, confirmatory factor analysis can be used to address the issues covered by multitrait-multimethod matrices. If the mix of tests in the analysis includes tests measuring different kinds of performance employing different methods, confirmatory models can be used to identify separate trait and method factors. Large loadings on trait factors and small loadings on method factors provide evidence for the convergence of the indicators for the traits. Large loadings on method factors would indicate a lack of convergence. Similarly, the trait measures are expected to load on the appropriate trait factor and not on factors associated with other traits; such results support the discriminability of the trait (McDonald, 1999; Messick, 1989).

3.7. Item Response Theory (IRT) Scaling

A large number of IRT models are available (see Yen and Fitzpatrick, this volume), each employing a particular mathematical expression to represent the probability that a person will get an item right as a function of the person's ability and the characteristics of the item, as represented by various item parameters. To the extent that the model holds, and the item parameters are known, we can predict how a person with a particular ability level will respond to any item or any set of items that fit the model.

The underlying ability scale (or theta scale) in an IRT model functions as a trait. It provides a summary of how well the person performs on the items used to define the ability scale.

> Fitting an IRT model is an empirical exercise, capturing and quantifying the patterns that some people tend to answer more items correctly than others, and some items tend to be answered correctly less often than others. The conception of ... competence embodied by the IRT model is simply the tendency to perform well in the domain of tasks. (Mislevy, 1996, p. 393)

Because IRT analyses are typically applied to the items included in a test or to all items used in a testing program, the resulting theta scale is tied to this pool of items.

The inference from a person's observed item performances to the person's ability level is analogous to an inference from an observed score to a universe score. Simple generalization relies on sampling models to warrant generalization to a universe score; IRT models warrant inferences from item performance to an ability level on the theta scale. Both models involve inferences from observed performances on some items to claims about overall performance on a universe of possible test items.

The observed performance is "explained or predicted by examinee characteristics referred to as traits or abilities" (Hambleton, 1989, p. 149), but no particular causal mechanism is specified. Nevertheless, the label assigned to the theta scale (e.g., inductive reasoning) may suggest the general nature of the processes involved in successful performance. In addition, the labels may imply that the ability is relevant to expected performance in certain contexts or situations (e.g., in an educational program or work environment). As is the case for any attribute, all of these implications are part of the interpretation.

In general then, IRT ability measures involve essentially the same four steps as the traditional trait interpretations. Item performances have to be scored. The results are extended to a claim about an underlying latent trait that is determined by the items used to define the ability scale, and by extension, is associated with a universe of possible test items that could have been used to define the scale. That is, the latent ability estimates serve a function similar to that of universe scores for traditional traits. The interpretation is then extrapolated to the target domain, and any implications associated with the attribute label are added. The main difference between the two approaches is in the second inference. In the classical approach, sampling models provide warrants for generalization. In IRT scaling, the IRT model provides a warrant for inferences from item responses to the estimated latent ability.

3.8. Trait Measures as Signs and Samples

Traits serve as inductive summaries over their target domains, which carry much of their meaning, but a trait interpretation also assumes that some underlying, latent attribute accounts for the observed regularities in the target domain. As Loevinger (1957) suggested, item responses are always both signs and samples of behavior; they are samples from the universe of generalization and from the target domain, and they are signs of the underlying attribute.

The interpretive argument for trait interpretations depends on substantive assumptions, statistical assumptions, and values. The scoring and extrapolation inferences depend mainly on the substantive and value assumptions implicit in the scoring rules and in the definition of the target domain. Substantive assumptions also play a large role in generating analytic evidence relevant to extrapolation and implication inferences. The statistical assumptions are embedded in sampling models (e.g., in generalizability analyses), scaling and equating models, and in the regression and other statistical models that can be used to examine extrapolation and various implications.

The interpretations and uses of test results are complex, and validation efforts should recognize this complexity. It is more important to state proposed interpretations and uses clearly and to evaluate them fully and critically than it is to have a tidy statistical analysis.

4. THEORY-BASED INTERPRETATIONS

The previous section examined attributes that derive most of their meaning from target domains. This section addresses attributes that are implicitly defined by their role in some theory. The validation of measures of these theoretical constructs necessarily involves an evaluation of the theory.

4.1. Theories, Constructs, Indicators, and Descriptive Attributes

Theory development begins with some phenomena to be explained by the theory. The phenomena are described in terms of existing observable attributes (or traits), which do not depend on the theory for their interpretation. These *descriptive attributes* are, in this sense, theory-neutral. For example, literacy can be defined in terms of a target domain of performances, without specifying any theory of performance, and therefore can function as a descriptive attribute in developing and testing theories of literacy.

The theory postulates some underlying mechanisms or relationships to account for the observed phenomena. In a quantitative theory, the postulates would be stated as equations relating the descriptive attributes to each other and to *theoretical constructs*, which represent aspects or components of the postulated mechanisms or relationships. The theoretical constructs are introduced to specify the theoretical mechanisms or relationships postulated by the theory and are implicitly defined by their roles in their defining theory.

As observable attributes, the descriptive attributes can be measured by drawing samples from their target domains (or more commonly, from standardized subsets of their target domains), and by using the results of these samples to estimate their target scores.

The theoretical constructs are not defined in terms of any domain of observations and therefore cannot be estimated so directly. The estimates of theoretical constructs generally rely on assumptions built into the theory. To take a very simple example, a theory of learning of some skill might assume that skill level (S) is directly proportional to instructional time (T) and to each person's aptitude (a_p):

$$S = a_p T$$

Skill level can be defined and measured as an observable attribute, and instructional time (T) is also observable. The aptitude, a_p, which is simply the slope in the hypothesized linear relationship, is not directly observable, but it can be estimated for a person by taking the ratio of the change in a person's skill level over some period of instruction to the length of instruction. The measurement of the aptitude depends on the theory.

An *indicator* of a theoretical construct is an observable attribute or a combination of observable attributes that is used to estimate the theoretical construct. The indicator does not define the construct, and a construct can have several distinct indicators. The indicator is necessarily based on observable attributes, but as an indicator, it is interpreted as an estimate of the construct defined by the theory. For this interpretation to be plausible, the theory must be plausible, and for the theory to be plausible, its predictions about observed relationships must hold. The simple theory sketched above predicts that skill level will be a linear function of the length of instruction and that the slope will be stable for each person but may vary from one person to another. The theory is evaluated by checking its predictions against data.

Descriptive attributes serve two important functions in the development and testing of theory. First, they specify the phenomena of interest. The phenomena have to be described and understood on some level before theories can be developed to explain them. Second, the descriptive attributes play a key role in evaluating the theory. They provide the "stubborn, dependable, replicable puzzles" (Cook & Campbell, 1979, p. 24) that theories are supposed to solve. If the theory's predictions about the descriptive attributes are accurate, the theory is supported, and if the predictions are inaccurate, at least some part of the theory is questionable.

If a theory is rejected, the interpretations of indicators in terms of the theory would also be rejected, but the interpretation of the descriptive attributes would not change much if at all. The failure of a theory to account for certain phenomena generally leads to a rejection of the theory and not a rejection of the phenomena.

4.1.1. Nomological Theories

Cronbach and Meehl (1955) framed their analysis of construct validity in terms of nomological theories, which specify networks of nomological (or "law-like") relationships among attributes. Structural equation models (SEMs) provide a modern statistical framework for analyzing such

nomological theories (Benson, 1998; Benson & Hagtvet, 1996; Jöreskog, 1973; McDonald, 1999). SEMs postulate causal relationships among latent variables and observable or manifest variables. The latent variables function as implicitly defined theoretical constructs that are linked to each other and to the manifest variables that function as descriptive attributes.

An explicit model is developed to explain the relationships among these latent variables and descriptive attributes, and the latent variables are estimated by fitting the model to the data. Empirical evaluations of how well the model fits the data serve as checks on the proposed model and as a basis for comparing alternative models of the same phenomena. These fit analyses also provide an empirical check on the validity of the indicators of the latent variables, which depend on the SEM for their meaning (Benson, 1998).

Cronbach and Meehl (1955) did not limit their discussion of nomological networks to well articulated statistical models. Rather, they suggested that

> the logic of construct validity is invoked whether the construct is highly systematized or loose, used in ramified theory or a few simple propositions ... We seek to specify how one is to defend a proposed interpretation of a test; *we are not recommending any one type of interpretation.* (p. 284) [emphasis in original]

They defined a construct broadly, as "some postulated attribute of people assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283), as in statements like, "Persons who possess this attribute will in situation X act in manner Y (with stated probability)" (p. 284). Given this loose conception of theories, some of the traits discussed in Section 3 can be considered constructs. As indicated in Section 1, Cronbach and Meehl's (1955) emphasis on stating the proposed interpretation clearly and evaluating it critically is generally applicable.

4.1.2. Process Models

Process models postulate specific cognitive processes to account for certain observed performances (Embretson, 1983; Ippel, 1986; Mislevy, Steinberg, & Almond, 2003; Pellegrino, Baxter, & Glaser, 1999; Sternberg, 1979; Tatsuoka, 1990). In addition to descriptive attributes specifying the performances of interest, a process model generally involves parameters (i.e., theoretical constructs) that characterize the latent abilities used to explain the observed performances (Embretson, 1983).

For example, a process model might assume that a number of component processes are called upon (sequentially or concurrently) to perform complex tasks. An IRT model can then provide estimates of latent ability parameters representing each student's level of skill in each of the component processes, as well as item parameters representing the level of processing difficulty of each task on each component (Embretson, 1984; Embretson & McCollam, 2000). The probability of success on a task is assumed to be the product of the probabilities of successfully completing the component processes involved in the task. The parameters serve as theoretical constructs, which are estimated by model-based indicators defined in terms of item responses.

In the context within which a process model is developed and empirically tested, the descriptive attributes describing performance on the tasks are taken as givens, and their definitions do not depend on the process model being proposed. In fact, alternative models could be developed to account for the task performances. A failure of any of these models would not generally have much impact on the measures of the descriptive attributes.

In contrast, the indicators of the model parameters depend on the model for almost all of their meaning. If the model were rejected, the indicators of the model parameters would go with it.

4.2. Interpretive Arguments for Indicators of Theoretical Constructs

A typical interpretive argument for indicators of theoretical constructs has five major inferences:

1. ***Scoring***: from observed performance to the indicator score
2. ***Generalization***: from indicator score to the universe score for the indicator
3. ***Extrapolation***: from the expected indicator score to the target score for the indicator
4. ***Theory-based Interpretation***: from the target score for the indicator to the construct as defined by the theory
5. ***Implications***: from the construct value to any implications suggested by the construct label or description

The first three inferences support an interpretation as a measure of the observable attribute (or combination of observable attributes) serving as an indicator of the construct. The fourth inference extends the interpretation to the theoretical construct, thereby expanding the interpretation to include all of the claims implicit in the theory. The fifth inference involves any additional implications associated with the construct label or description.

4.3. Validity Arguments for Indicators of Theoretical Constructs

The first step in validating any proposed interpretation is to evaluate the coherence and completeness of the proposed interpretive argument. If the structure of the interpretive argument is considered satisfactory, the specific inferences and assumptions in the argument are evaluated.

The first three inferences in the interpretive argument (scoring, generalization, and extrapolation) serve to define the indicator as an observable attribute, but the scoring inference may be quite complex for indicators of theoretical constructs. In particular, the scoring inference may require that various performances be scored and that the resulting scores be combined to yield the indicator. If the constructs are parameters in IRT models, the scoring inference tends to be relatively straightforward, but the generalization inference is replaced by parameter estimation (see Yen and Fitzpatrick, this volume), and the extrapolation inference is subsumed under the theory-based interpretation.

The interpretation of indicator scores as estimates of a theoretical construct extends the interpretation to a claim about a construct as defined by the theory. Theory-based interpretations of indicator scores assume that the theory provides a sound explanation for the relevant phenomena and that the indicators provide appropriate estimates of the constructs in the theory. The warrant for this inference is the theory. The backing for the warrant would involve analytic and empirical evidence supporting the theory and the appropriateness of the indicator. The analytic evidence relies on analyses of the relevance of each indicator to its construct and is produced during the development stage, as the indicators are designed to fit the theory. The empirical evidence examines how well the theory's predictions agree with observable phenomena.

4.3.1. Analytic Evidence for Theory-Based Inferences

Initial confidence in the theory-based inferences depends in part on judgments about how well the indicator represents the construct as defined by the theory. A theory provides guidance on how to develop indicators for its constructs. If the theory predicts that individuals with high values on a construct are likely to perform well on some tasks and that individuals with low values of the construct are likely to perform poorly on the tasks, performance on the tasks can be used to estimate the value of the construct.

The theory may also indicate various conditions of observation that could have a substantial impact on the observations. In developing an indicator, efforts would be made to control any potential source of irrelevant variation. Basically, indicators are designed to provide plausible estimates of the construct and to be relatively free of systematic and random error.

In general, it is desirable that the indicator involve a broad sampling of relevant observations (Cook & Campbell, 1979; Messick, 1989). By including a broad range of observations associated with the construct, the possibility of construct under-representation is minimized, and by including multiple modes of observation, construct-irrelevant variance can be controlled. If serious efforts are made to identify potential sources of irrelevant variance, and no likely candidates are identified, the interpretation is supported. If any potentially serious sources of systematic error are identified, the accuracy of the indicator as an estimate of the construct is questionable.

4.3.2. Critical Appraisal of Theory-Based Inferences

Most of the evidence needed to evaluate a theory-based inference is that needed to evaluate the theory. The theory is evaluated in terms of its general plausibility and by subjecting it to empirical challenges. A theory that survives serious challenges is accepted as a working assumption (Cronbach, 1980b; Lakatos, 1970; Popper, 1962).

Any prediction about observable attributes that can be made from the theory provides a potential empirical check on the theory. These empirical checks include predictions about performance or behavior in various contexts or by various groups, correlational evidence, and the results of experimental studies. To the extent that the predictions are verified, the plausibility of the theory is enhanced, and therefore, the plausibility of the proposed theory-based interpretation is enhanced.

If the results do not agree with the predictions, some part of the theory is called into question. One option is to reject the theory as a whole. A second option is to reject one or more of the indicators. A third option is to assume that some external factor distorted the results. For any particular study, it is generally not obvious which of these three options would be appropriate, and therefore, no single study will be decisive. The plausibility of the theory and of interpretations based on the theory will depend on the results of a number of studies.

4.3.3. Correlational Analyses

As noted in Section 2.3, correlation coefficients provide much of the empirical evidence produced during test development. In particular, evidence for the homogeneity, or internal consistency, of the observations in the universe of generalization supports generalizability over tasks, and correlations between test performance and non-test performance can support extrapolation from the universe score to the target score for the indicator. When developing indicators of several distinct constructs at the same time or when introducing a new construct indicator, the consistency among different indicators of each construct, and the ability of the indicators to distinguish among different constructs, can be evaluated using multitrait-multimethod matrices or confirmatory factor analysis.

At the appraisal stage, correlational evidence is particularly relevant to the evaluation of nomological theories. Nomological theories predict certain kinds of relationships (e.g., strong or weak, direct or inverse) among observable attributes (descriptive attributes and indicators). If the number of relationships is small or the expected pattern is relatively simple, it may be possible to evaluate the fit between the predicted relationships and the observed relationships directly. More complicated structural, causal, and hierarchical hypotheses can be investigated using relatively general models like structural equation models, confirmatory factor analysis, and hierarchical linear models.

The role of these analyses in the appraisal stage of validation is to subject the proposed theoretical interpretation to empirical challenge. If one theory predicts a positive correlation between two variables and a second theory predicts a negative correlation under the same circumstances, empirical results may support one theory over the other.

4.3.4. Experimental Manipulation

If the theory hypothesizes certain causal relationships, such that a change in one variable is expected to produce some change in a second variable, then it may be possible to check the theory by manipulating the first variable and monitoring the second variable. If the expected changes occur, the theory is supported; if not, the theory tends to be refuted.

For example, some theories of anxiety indicate that motivation to avoid failure would tend to promote test anxiety

that would in turn have a negative effect on achievement (Benson, 1998). So, by reducing the need to avoid failure in some way (e.g., by reducing the stakes inherent in testing) it should be possible to reduce test anxiety and thereby to improve achievement. If this pattern were observed, the theory would be more credible; if not, the theory would be less credible.

For process models, it may be possible to improve skill on certain component processes by focusing instruction on the processes. Such focused instruction, if effective, should increase level of skill on that process, and as a result, improve performance on tasks that require that component. Note that the conclusion contains the caveat that the instruction must be effective, and therefore, a finding that instruction has not produced the expected change may not produce a decisive refutation of the theory.

Some attributes are not expected to be affected much by short periods of intense instruction. General abilities (e.g., general verbal or quantitative reasoning) are expected to be less susceptible to targeted instruction than specific skills (e.g., solving algebraic equations). Therefore, if scores on a test of general ability improved substantially after a brief coaching program, the validity of the test scores as indicators of the general ability would be questionable (Messick, 1989).

4.3.5. Evaluating the Backing for Theory-Based Inferences

The overall plausibility of the theory and of theory-based inferences depends on all of the evidence, for and against the theory. The strength of the backing for a theory-based warrant depends on a number of factors, including the number and range of predictions that are examined, the uniqueness or novelty of the predictions, and the quality of each empirical study. Surviving one kind of empirical challenge, even if this challenge has been examined in a number of studies, is not as convincing as surviving a range of different challenges in different contexts (Loevinger, 1957).

Disagreements between the predictions and observations always count against a theory. Agreement between predictions and observations provides strong support only if the predictions cannot be taken for granted in advance. For example, a prediction that performance on some task will generally improve with practice does not provide strong support for a theory of task performance, because this relationship holds for most tasks in most contexts. Agreement between theory and observation does not provide support for one theory relative to another, if both theories make the same prediction.

The impact of any particular study also depends on the quality of the study. A large, well designed study that has eliminated the most serious threats to its conclusions provides more powerful evidence than a smaller study that is subject to various sources of systematic error. For example, finding a predicted positive relationship between two variables that are assessed with the same method (e.g., a rating scale) or the same raters provides weaker confirmation than would be the case if the same correlation resulted from observations involving different raters, using different methods, on different occasions. The latter findings provide stronger evidence because they are less likely to be due to method artifacts.

4.3.6. Challenges to Indicators of Theoretical Constructs

On a fundamental level, theory-based construct interpretations can be challenged by claiming that the theory is not plausible. If the theory cannot stand up to criticism, interpretations based on the theory are not defensible.

Even if the theory as a whole is considered credible, the appropriateness of a particular indicator for a theoretical construct can still be challenged in terms of either construct under-representation or construct-irrelevant variance (Cook & Campbell, 1979; Messick, 1989). For indicators of theoretical constructs, questions about representativeness focus on the extent to which the observations defining the indicator elicit the full range of processes associated with the construct (Bachman, 2002; Borsboom, Mellenbergh, & van Heerden, 2004; Embretson, 1983; Loevinger, 1957). Any factor other than the construct that has an impact on the indicator (e.g., the method or context of observation) is a source of construct-irrelevant variance, or systematic error.

Some systematic errors apply generally in the population and constitute general threats to validity. Some systematic errors may have a particular impact on certain individuals or groups. Toulmin (1953) has argued that scientific laws or theories are not true or false, but that "statements about their range of application can be" (p. 79). If an otherwise successful theory fails to account for certain observations, it may be decided that the theory simply does not apply under certain circumstances. Under this view, the issue in empirically testing a theory is not to find out if it is true or not, but to find out how widely it applies. Exceptions to model-based inferences are triggered by indications that the model does not apply in a particular case.

More general exceptions occur if there is evidence suggesting that the theory does not apply to particular groups, or if almost all or all of the evidence supporting the model or supporting the connection between assessment scores and model parameters was developed for some groups, and there is reason to believe that the model might not apply to other groups. For example, a performance model might have been developed and shown to apply for students in instructional programs based on the model (e.g., with topics covered in a particular sequence), but the performance model might not apply so well to students in other instructional programs. Evidence indicating that the performance of students at one grade level conforms to a particular model does not ensure that the performance of students at a different grade level will conform to the same model. This is not to say that a theory cannot be extended beyond the contexts in which it has been empirically checked, but rather, to suggest that such extensions involve risk.

4.3.7. Evidence for Construct Implications

In addition to the basic interpretation of a construct as defined by its theory, the label or uses of the indicator scores

may suggest additional claims. The adoption of a theory of intelligence would define the meaning of intelligence as a theoretical construct within the theory, but claims about the relevance of this construct to success in non-test contexts would typically go beyond the theory.

Embretson (1983, 1998) distinguishes two aspects of validity; construct representation and nomothetic span. *Construct representation* is concerned with "the processes, strategies, and knowledge that persons use to solve items" (Embretson, 1998, p. 382). It is supported by evidence that supports the proposed cognitive models. Embretson (1983, 1998) contrasts the meaning of the construct embodied in the explanatory model from its *nomothetic span*, which describes the relationships of test scores with other measures. Nomothetic span emphasizes significance or utility rather than meaning. Within the framework adopted here, construct representation can be associated with the meaning of the construct, while nomothetic span describes additional implications associated with the construct label.

The validation of a proposed interpretation or use of test scores requires an evaluation of all of the claims inherent in that interpretation or use. For example, the use of aptitude test scores for placement into alternative treatments depends on the assumption that the regressions of achievement on aptitude for the different treatments will cross each other (Cronbach & Snow, 1977). Whether this requirement is satisfied by a particular aptitude test in a particular setting is an empirical question that may transcend the theoretical basis for the aptitude measures.

4.4. The Role of Theories in Interpreting Descriptive Attributes

While theory-based explanations are not necessary for the validation of measures of descriptive attributes, such explanations can provide strong support for some of the inferences in the interpretive arguments for these attributes. The theory may provide explicit guidance for test development and justification for test content, format, and scoring rules by indicating the skills that are most strongly associated with performance on the descriptive attribute.

The theory can also provide support for generalization by indicating the task characteristics or conditions of observation that are likely to have large impacts on performance. For example, the representativeness of a sample of subtraction problems is more plausible if it can be shown that it includes problems requiring all component skills involved in subtraction. To the extent that a model can accurately predict the difficulty of different tasks, the model can be used to control test difficulty by making sure that all persons get tests of the same difficulty or by adjusting the scoring rule for each test to reflect differences in difficulty (e.g., using equating).

To the extent that the model is known or expected to apply in varied contexts (e.g., in the classroom, at work, in the community), it can also provide a solid basis for extrapolating from test performance to performance on other kinds of tasks in other settings (e.g., the workplace). The model thereby provides additional support for extrapolation by providing a rational basis for this inference.

The descriptive attribute does not rely on the theory for its meaning. However, descriptive attributes that are attached to well-confirmed theories can be used to draw conclusions and make predictions based on the theory. The theory thereby adds to the implications or nomothetic span of the descriptive attribute.

4.5. Construct Validity and Theory Testing

In the traditional view of science, the theory's predictions are compared to observations, which "serve as the universal, neutral arbiter among alternative hypotheses" (Galison, 1987, p. 7). The observations are taken as given, and the theory is evaluated by how well it predicts (or explains) the observations (Lakatos, 1970; Popper, 1962). If the theoretical predictions agree with the observations, confidence in the theory increases, and if not, confidence in the theory decreases.

However, in Cronbach and Meehl's (1955) formulation of construct validity, the theory is effectively taken as given, and the appropriateness of the indicators is evaluated by how well the theory fits the data. The evaluation of theories becomes quite complicated if the interpretations of measurements depend on the theory and, therefore, are as much in doubt as the theory (Quine, 1953). A system in which theories are evaluated by comparing their predictions to measurements, and the measurements are validated in terms of the theory, clearly has the potential for circularity. Loevinger (1957) expressed concern that Cronbach and Meehl's use of theories to define constructs could be taken to imply that "validation studies are communicable only among such coteries as are agreed on theoretical issues" (p. 643).

In practice, the distinction between descriptive attributes and indicators of theoretical constructs can be used to cut the circle and make both theory testing and test validation manageable. Theories are developed within a scientific community to explain some phenomena, and the phenomena to be explained can be specified in terms of descriptive attributes without relying on the theory (Galison, 1987; Guion, 1977). The predictions made by the theory about these descriptive attributes provide empirical checks on the theory as a whole, including the choice of indicators for theoretical constructs (Cronbach & Meehl, 1955). That is, the descriptive attributes are taken as given, and the theory is evaluated as a whole by comparing its predictions about the descriptive attributes to the measures of these attributes.

Empirical evaluations of competing theories are possible, because a community can agree on the definition of relevant descriptive attributes and on acceptable measures of these attributes. For example, it is possible to define and measure achievement in arithmetic in terms of a target domain of performances, without developing a theory that explains how schools teach arithmetic or a model that explains how people solve arithmetic problems. The interpretation of the descriptive attribute does depend on various ancillary assumptions, (e.g., about what constitutes an adequate performance on a task) but, once these specifications are agreed on, the measure can be taken as an observable attribute and can serve as a neutral arbiter in evaluating competing theories (Grandy, 1992).

5. QUALITATIVE INTERPRETATIONS

Most decisions in classrooms, clinics, and other real-world settings are based on qualitative assessments, which focus on the interpretation of performance in context and give little or no attention to scores. Qualitative assessments are reported as narrative interpretations, or "thick descriptions" (Geertz, 1973), rather than as interpreted scores. The teacher/clinician collects information and acts on it at the same time, seeking to develop a coherent interpretation that is consistent with all available evidence (Moss, 1994; Shepard, 2001). Classroom assessment of student performance will serve as the primary focus of this section (Frederiksen, 2003; Gipps, 1999; Moss, 1994; Moss & Shutz, 2001; Shepard, 2001, this volume; Stiggins, 2005; Tittle, 1989).

For classroom and other clinical assessments, the goal is not to subsume observations under empirical laws, but, "to place them within an intelligible frame" (Geertz, 1973, p. 26). Different kinds of information from different sources are combined for an interpretation of performance in context (e.g., that of a student in a class). Instead of the observation-scoring-interpretation paradigm prevalent in standardized testing, qualitative assessment involves an active search for meaning from the beginning, with the interpretation being elaborated and extended as data are collected. The goal is to construct a coherent interpretation of performance, "continually revising initial interpretations until they account for all of the available evidence" (Moss, 1994, p. 8). The qualitative approach focuses on evolving interpretations of observations rather than on scores.

Unlike quantitative interpretations, which suppress many interactions among variables by standardizing the measurement procedure and by averaging over any variables that are considered irrelevant, qualitative assessments attend to interactions among the factors that influence performance. As a result, they can provide a richer interpretation of events in environments where interactions among causal factors are the norm (Cronbach, 1975).

5.1. Qualitative Interpretations of Student Performance

Teachers have access to a rich array of data that can be helpful in developing their understanding of each student (Brookhart, 2003; Moss, 2003, Stiggins, 2005; Tittle, 1989). They observe their students' performance on a variety of educational tasks (e.g., participating in classroom activities and discussions, taking tests, completing projects) over an extended period. They talk to the students, to their parents, and to other staff. Teachers may also have scores on standardized tests and information on student's interests, health (including any disabilities), and backgrounds.

Teachers have to make sense of all of this data, and in doing so, they employ an array of conceptual resources and frameworks to organize their observations (Shepard, 2001). They understand the structure of what is being taught and have goals for what students are to achieve. In addition to content-specific skills, most instruction seeks to foster the development of various general skills (reasoning, problem solving, creative expression, reading, writing, mathematical skills), and teachers track each student's progress in developing these skills.

For example, if a second-grade teacher observes a student reading a story aloud but stumbling over some more difficult words, the teacher may conclude that the student can read, but that his or her vocabulary is being stretched by the story. The performance is interpreted in terms of the teacher's conceptual frameworks (e.g., their understanding of the processes involved in reading). To the extent that the teacher is familiar with the student and with the difficulty level of the text, a richer explanation is possible.

Teachers typically have a sense of the kinds of student misconceptions and use this understanding to explain student errors. They also have some sense of the social and organizational context and the expectations and limitations inherent in these contexts and are familiar with various resources (e.g., texts at different levels of difficulty, reference materials) that they and students can make use of in teaching and learning.

Teachers bring a potentially rich set of tools to their observations of student performances. Experienced teachers meeting their new class for the first time have a good sense of the range of skill levels to expect. They know some of the gaps in understanding and skill to expect, and common impediments to learning, and are familiar with "the relevant experience and discourse patterns" (Shepard, 2001, p. 1075) of their students. Their conceptual frameworks provide templates for organizing observations of student performance and for differentiating critical issues from more routine observations. Using these frameworks, teachers interpret student performances as they occur and do not simply keep a record of their observations for later interpretation.

5.1.2. Refinement of Teachers' Views of Their Students

A teacher's understanding of what students know and can do, of their aptitudes, limitations, and interests evolves as the teacher interacts with the students. Initially, the teacher's understanding of each student may be quite general, but over time, it is developed and refined as the teacher interacts with the student. For example, if a student who has shown little interest in reading now exhibits a strong interest in Sherlock Holmes stories, the teacher's view would be modified to reflect this development, and instructional choices could reflect this more nuanced description. The teacher's evolving view of the student is neither complete nor completely accurate, but it is likely to get better over time.

Teachers use their evolving views of the students to guide their interactions with students in various contexts. These views generate expectations about student performances on various tasks in various contexts. The teacher does not generally predict future events but does anticipate them (Geertz, 1973) in the sense that, for a particular student and situation, some kinds of events are seen as more likely than others. If these expectations are confirmed, the teacher's confidence in his or her current views increases. To the extent that the expectations are not confirmed, the teacher may modify assumptions about the student, the tasks, or the context. The teacher's view of each student develops over an extended period and can be self-correcting (Black & William, 1998; Delandshere, 2002; Lane, 2004; Moss, 2003; Stiggins, 2005).

If a teacher's expectations fail for many students in many situations, the teacher may need to rethink some of their basic assumptions. On the other hand, if the teacher's conceptual frameworks are working reasonably well for most students but are failing for a particular student, the teacher is more likely to rethink the assumptions being made about that student. For example, a teacher would use his or her current view of a student in interpreting the performance of the student in reading a passage aloud and answering questions about the passage. Given the difficulty levels of the passages that the student has read with good comprehension and those that seem to be beyond the student, the teacher has expectations about the student's performance on a new book. If a student has trouble with a book that the teacher expected the student to be able to read, the teacher may conclude that the student is a weaker reader than the teacher had thought, that the book is more difficult than the teacher thought, or that the context or some extraneous effect interfered with the student's performance.

Note the lack of symmetry here. If the student reads the book, it is reasonable to conclude that the student has the skills needed to read the book. If the student does not read the book, there are multiple possible explanations for the event, involving the student, the book, or the context. If the student had previously read more difficult books, the teacher may be inclined to attribute the poor performance to something other than a lack of skill in reading (e.g., lack of interest in the book's content, some aspect of the context). If the book is more challenging than those the student had been reading, the student's poor performance could reasonably be attributed to the student's reading level relative to the difficulty of the book. The teacher may adopt (perhaps tentatively) one of these explanations or may defer the choice until more information is available.

The teacher's view of a student may involve some definite expectations (e.g., that a student will be able to perform certain tasks) or some range of possibilities (e.g., that the student may be interested in certain activities). For observations that share many features with those on which the evolving view is based, the expectations may be expressed with some confidence. For new observations that involve many new elements (a substantially different task or setting), the expectations would generally be more tentative.

5.1.3. Interpretive Arguments for Classroom Assessments

The interpretations of teacher assessments of their students can be represented by an iterative interpretive argument:

1. ***Development of Initial Views of Students***: Teachers use their conceptual frameworks to develop a view of the student. They use these frameworks to make sense of their initial interactions with students and to develop their initial views of students. The conceptual frameworks provide a basis for evaluating student performance and for anticipating various factors that might influence student learning and performance.
2. ***Refinement of the Teacher's View***: Teachers modify their views of students, based on their ongoing interactions with the students. The teachers' evolving views generate expectations (e.g., that a student will or will not be able to read certain books or to solve certain kinds of problems). These expectations provide working assumptions for subsequent instruction and assessment, and comparisons of the expectations to subsequent observations provides feedback on the accuracy of the teacher's views.
3. ***Extension of the Teacher's Evolving View to New Contexts***: Teachers may also use their overall assessments of students to draw conclusions about expected performance in new contexts (e.g., beyond the classroom). At various times and for various reasons, the teacher may need to draw general conclusions about student progress (e.g., in making end-of-year promotion decisions).

5.1.4. Teachers' Conceptual Frameworks as Theories

This kind of reasoning represents an informal version of theory testing in science (Cronbach, 1975; Toulmin, 2001), with the teachers' conceptual frameworks and their evolving views of students functioning as theories that generate expectations about what is more or less likely to happen in various situations. If these expectations are confirmed, confidence in the frameworks and evolving views is enhanced. If the expectations are contradicted, changes may be called for in some part of the teacher's views.

Qualitative interpretations are also analogous to the more formal construct interpretations discussed in the last section. In both cases, a general model, or "working hypothesis," provides explanations of performance (Cronbach, 1975), but there are some clear differences between the two approaches. In particular, cognitive-process models assume that the same model applies to all test takers, and second, the assessment procedures based on these models tend to be highly standardized. In such model-based assessment, the goal is to estimate the parameters in the model for each individual. In classroom assessment, the teacher integrates data from a variety of sources, most of which are not standardized, to develop a holistic view, or model, of each student.

5.2. Validity Argument for Classroom Assessments

The validity argument begins with an evaluation of the completeness and coherence of the interpretive argument and an evaluation of whether the interpretive argument provides a reasonable explication of the proposed interpretation. If the general form of the interpretive argument is satisfactory, its inferences and assumptions would then be evaluated. Taking the interpretive argument developed above as a working model, the validity argument would focus on the plausibility of the teacher's conceptual frameworks and of their evolving views of students.

5.2.1. Development of the Teacher's Views

The teacher's conceptual framework cannot generally predict what is going to happen, but it does allow the teacher to anticipate various possibilities, and it can explain observations after they have occurred. The teachers' conceptual

frameworks provide the warrants for these interpretive evaluations. The backing for these warrants includes the backing for the teachers' conceptual frameworks (e.g., conceptions and organization of subject matter, pedagogical theories and techniques) and the evidence supporting the ability of the teacher to use these tools effectively, including the training, credentials, and experience of the teacher, the extent to which the teachers have access to relevant data, and any quality-control safeguards that are in place.

Teaching involves an ongoing stream of interactions and decisions, and most of the backing for a teacher's conclusions about students is not publicly available. Teachers do not record all of the reasoning involved in their conclusions, nor do they specify the inferences leading to most of their decisions. However, if asked to do so, a teacher could presumably provide a rationale for his or her choices.

Additional support for the teacher's inferences can be provided by peer or external review. Moss (1994) emphasizes the role of dialogue within a "critical community" of individuals with expertise on the issues involved, as well as those with a stake in the outcome (e.g., parents, community members, students). Dialogue-based warrants are likely to work well in cohesive, autonomous communities with a base of shared assumptions. Within this context, disparate opinions serve several important functions; in particular, they can identify weaknesses in the proposed interpretation and can illuminate implicit assumptions and potential biases (Moss, 1994; Ryan, 2002).

5.2.2. Refinement of the Teacher's Views

Teachers' evolving views of students, supported by their conceptual frameworks, provide the warrants for their ongoing interpretations of their interactions with students. The backing for these warrants is provided mainly by evidence supporting the adequacy of the teacher's evolving views, as a basis for making sense of their interactions with students.

Teachers do not generally rely on statistical inferences in interpreting student performances, but rather on less formal presumptive reasoning. An evaluation of a single performance by a single teacher can justify part of an evolving view. If a qualified teacher evaluates a student's performance on some complex activity (e.g., solving a quadratic equation, making a bowl on a potter's wheel) and uses these observations to draw conclusions about how well a student has mastered the skills needed to perform the activity (e.g., solve equations, make bowls), it is reasonable to accept the teacher's judgment at face value, unless there is some specific reason to question it. As Frederiksen (2003) has suggested, a qualitative analysis of performance

> may reveal that a student has used problem-solving approaches and forms of knowledge that are highly generalizable to other task situations, thus backing a student-model claim about generality of skill. (p. 71)

With the help of their general assumptions about student performance (based on their understanding of task requirements and experience with students) teachers can derive general conclusions about student competencies from samples of performance.

This kind of inference is widely used in evaluating scientific theories and can support strong conclusions. The underlying rationale is analogous to that for Bayes' Theorem; if the probability of an event is high given some hypothesis and is low under all alternative assumptions, then the occurrence of the event provides strong evidence for the hypothesis. For example, given that there is a high probability of getting the correct solution to a quadratic equation if one knows how to solve such equations, and a very low probability of success if one does not know how to solve such equations, students who generate correct solutions to a quadratic equation that they have not seen before can be assumed to know how to solve quadratic equations. In addition, if the teacher asks the student to explain the performance and the student provides an accurate, coherent explanation, it is reasonable for the teacher to conclude that the student can solve quadratic equations.

By examining a body of student work (e.g., in class or in a portfolio), the teacher can form hypotheses about the student's competencies and about gaps in the student's understanding of a topic. If a particular set of conjectures about a student does account for the student's pattern of performance (including the mistakes), and no plausible alternative hypothesis does as well, the proposed conjectures can be accepted as a reasonable conclusion about the student.

Since the qualitative interpretations do not necessarily involve statistical generalization (i.e., from a sample of observations to the expected value over a universe of generalization), it is not necessary to make a case for this kind of generalization (Moss, 1994), but it is necessary to provide support for the inferences from specific observations to a general description of the student. These presumptive inferences are based on content and context-specific warrants. The conclusion about a student's ability to solve quadratic equations rests on our understanding of how such equations can be solved and on experience with students trying to solve them. For content that is less well structured, this kind of model-based inference from performance to competence tends to be less precise. The fact that a student can name the capitol of one state provides little assurance that he or she knows any other state capitols.

If a qualified teacher applies appropriate criteria to a student's performance, the results would have a strong presumptive claim on our confidence. This is the kind of evaluation that might well serve as a criterion in validating some standardized test.

Like all measurements, teacher assessments of their students can be challenged. Even if a qualified, experienced teacher evaluates a student's performance over a long period, a challenge could claim that there was some mistake or omission in the teacher's assessments.

In practice, requests for backing for any particular aspect of a teacher's views of their students may come from the student, a parent, a fellow teacher, or a principal, and the teacher would be expected to respond with an explanation. This kind of dialogue encourages the teacher to articulate and evaluate their own views in a critical way.

5.2.3. Extension of the Teacher's Views to New Contexts

Once a teacher's evolving views of their students have been developed and refined, conclusions based on these views may be extended beyond the particular classroom (e.g., in evaluating a student's readiness for the next grade). The teacher may reasonably draw conclusions about expected performance in settings and on tasks to which student skills would be expected to transfer (based on the teacher's experience, and their understanding of the tasks and situations). As Frederiksen (2003) pointed out, some problem-solving approaches and skills are highly generalizable.

Although formal empirical studies are generally not feasible in evaluating classroom assessments, some kinds of teacher expectations can be compared to subsequent outcomes. For example, it may be possible to examine the subsequent performance of students whom the teacher recommends for promotion or retention.

These extensions of the teacher's conclusions about students to contexts beyond the classroom may be higher stakes than classroom assessments, because they may not be easily correctable (e.g., promotion decisions). The rationale for these extensions is likely to be stated explicitly and in some detail with supporting evidence.

Like all presumptive inferences, teacher assessments are defeasible and can be challenged on various grounds. Information about the student or the environment or the interaction between the two can suggest the need to revise an interpretation. For example, if a teacher had not taken certain information about a student (e.g., limited English language proficiency or a disability) into account in a context in which it was relevant, the interpretation could be questioned. And just as teacher judgments can support challenges to test scores for individual students, test scores can support challenges to teacher judgments.

5.3. Qualitative Approaches to High-Stakes Assessments

As the stakes associated with an assessment go up, the form of the assessment and the shape of the interpretive argument do not necessarily change, but the need to document the procedures being used and to provide backing for the warrants being applied increases. In cases where the proposed interpretation is extended beyond the local classroom setting, it is generally necessary to document the evolving description of each student more thoroughly than would be necessary within a single classroom with a single teacher. In these higher-stakes contexts, the teacher's evolving views may need to explicitly incorporate examples of student work so that readers "may judge its adequacy for themselves in supporting the desired generalization" (Moss, 1994, p. 8).

As the stakes go up, there is also pressure to increase standardization (e.g., by specifying the format and general content of student portfolios) in order to promote comparability of conclusions across settings and occasions, and thereby, to promote a kind of objectivity (i.e., a lack of subjective judgment) that enhances credibility for many audiences (Porter, 2003). As Shepard (2001) has noted, standardization "involves a basic matter of fairness" (p. 1081). Standardization also facilitates aggregation of results to the school, district, and state level.

In high-stakes contexts, where a range of stakeholders have a strong interest in the outcomes, it is generally expected that the analysis of student performance will be documented in a written report, which provides examples of student work and the rationale for the conclusions being drawn. The report would include warrants for major inferences and backing for any warrants that might be questioned by the intended audience. Employing multiple sources of evidence and input from independent observers strengthens the case. Tone and substance that suggest the absence of bias is especially important:

> The vigorous attempt to discover problems with the proposed interpretation—the search for disconfirmatory evidence and for alternative interpretations that account for the same evidence—is central to the development of well warranted interpretations. (Moss, 2003, p. 18)

Contradictory evidence does not necessarily disconfirm the proposed interpretation, but it does introduce notes of caution and can help the reader of the report to form a more complete view of the student.

For assessments that are used within a classroom, consistency in standards across classrooms is largely irrelevant. In fact, it is clearly appropriate that the teacher's assessments be tailored to the needs of his or her students. However, if interpretations and decisions are extended beyond individual classrooms (e.g., in statewide assessment programs) inconsistencies across classes and schools can cause problems.

Empirical research indicates that assessment standards vary across teachers and schools. Essay questions are more standardized than teacher observations, but essay graders generally find it difficult to maintain consistent standards over time. Studies of the consistency in the scores assigned to portfolios are not encouraging (Klein, McCaffrey, Stecher, & Koretz, 1995; Nystrand, Cohen, & Dowling, 1993). In a comparison of standards across states, Linn, Kiplinger, Chapman, and LeMahieu (1992) found that although there was a high consensus about the characteristics of good writing, as indicated by high correlations among graders from different states, there was "substantial diversity in the implicit standards of raters from different states" (p. 104). It is not easy to maintain consistent standards in evaluating complex performances.

Quality-control procedures seem to be the best option for dealing with concerns about the consistency of standards. For example, *social moderation* (or verification) procedures in which teachers' grades on samples of student work are compared to independent ratings by other teachers from the same or different schools and by external expert raters can build confidence in the consistency of standards across classrooms, schools, districts, and states (Linn, 1993). Social moderation aims to build consensus across schools:

> The process of verification of a sample of student papers or other products at successively higher levels in the system

(e.g., school, district, state, or nation) provides a means of broadening the consensus across the boundaries of individual classrooms or schools. It also serves an audit function that is likely to be an essential element in gaining public acceptance. (Linn, 1993, p. 99)

For large-scale, high-stakes assessments (e.g., those used for accountability), the audit function is essential.

5.3.1. The Role of Interactions in Qualitative Interpretations

In general, teachers use their conceptual frameworks and experience to make sense of their encounters with their students and of the students' work. The teacher does not start with an operationally defined scoring inference and then seek to interpret this score. The interpretation is built in from the beginning.

Rather than generalizing from a sample of performances to the expected level of performance over some universe of exchangeable performances, the teacher seeks to integrate all available evidence on each student into a coherent view of the student. The goal is to develop a narrative that makes sense of the full body of available evidence, and the proposed interpretation is refined through an iterative process of developing an interpretation and then checking it against further observations (Tittle, 1989).

In contrast, quantitative models seek to minimize the impact of conditions of observation, of the social and physical context, and of extraneous student characteristics (e.g., motivation, anxiety) by averaging over conditions of observation or by standardizing some conditions of observation (Cronbach, 1975). For many purposes, the use of statistical generalization and extrapolation to average over most interactions is an effective strategy. In most classroom and clinical contexts, the interactions merit attention.

6. DECISIONS

The use of tests to make decisions introduces some new wrinkles into validation, but the same basic approach still applies. The intended interpretations and uses are specified and are evaluated by examining their coherence and the plausibility of their inferences and assumptions.

Test-based decision procedures typically incorporate an interpretation in terms of attributes (traits, constructs, or qualitative descriptions) followed by a decision based on the interpretation. Both the substantive interpretation and the decision rule indicating what to do under various circumstances are included in the rationale for test use.

The addition of a decision rule to an interpretation introduces (or at least brings to the foreground) the issue of consequences. While the interpretations discussed in Sections 3 to 5 are evaluated in terms of their coherence and plausibility, decisions are evaluated in terms of their outcomes, or consequences.

This section has three parts. Section 6.1 discusses general characteristics of test-based decision procedures and their validation, defines some terminology, and raises some salient questions. Section 6.2 analyzes the validation of school accountability programs as an example of a test-based decision program. Section 6.3 addresses the role of consequences in validation (particularly social consequences) a topic that has been somewhat contentious (Guion, 1995; Messick, 1989; Popham, 1997; Sackett, 1998; Shepard, 1997).

6.1. Semantic Interpretations and Decisions

For purposes of this discussion, it will be useful to distinguish between semantic interpretations and decisions. A *semantic interpretation* draws conclusions based on assessment results (e.g., about a trait or construct), and thereby, assigns meaning to these results. The inferences and assumptions involved in the semantic interpretation will be referred to as *semantic inferences* and *semantic assumptions* respectively. A semantic interpretation is evaluated in terms of its plausibility.

A decision procedure implements a policy and involves choices about what to do; it is evaluated in terms of its outcomes, or consequences. The assumptions supporting a decision procedure make claims about the outcomes likely to result from the policy and about the values associated with these outcomes. In particular, they claim that the decisions will generally have positive consequences (e.g., placement of students in appropriate classes) and that any negative consequences will not be too serious. Overall, it is assumed that the positive consequences will outweigh the negative consequences.

Evaluation of a decision procedure necessarily involves an evaluation of values and consequences. Policies are not true or untrue, accurate or inaccurate. They are effective or ineffective, successful or unsuccessful. A policy that achieves its intended goals (positive consequences) at modest cost, and with few undesirable side effects (negative consequences) is considered a success. A policy that does not achieve its goals (lack of positive consequences), and/or that involves relatively high cost or produces significant undesirable side effects (negative consequences) is considered a failure.

For many test-based decisions, the semantic interpretation and the decision are distinct and sequential. For example, a course-placement test typically has a semantic interpretation in terms of level of achievement, and the placement decision is based on this semantic interpretation. The semantic interpretation and the decision may be made by different individuals at different times and in different places (e.g., college admission test scores reported by a testing agency later used by colleges to make admissions decisions). In other cases (e.g., classroom testing), semantic interpretations and decisions are intertwined.

The role of validation in evaluating the plausibility of the semantic interpretation and its role in evaluating the legitimacy of the decision procedure are both recognized in the *Standards*:

> Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. (AERA et al., 1999, p. 9)

The *Standards* go on to say that validation involves the development of evidence, "to support the intended interpretation of

test scores and their relevance to the proposed use" (AERA et al., 1999, p. 9). Although they have differed somewhat in emphasis, both Cronbach (1971, 1980a) and Messick (1975, 1980, 1989), have associated both interpretive accuracy and consequences with validity (Moss, 1992).

In principle, the evaluation of the effectiveness of a decision rule involves an application of utility theory (Cronbach & Gleser, 1965). In practice, the analysis is usually qualitative, because it is generally not feasible to specify the required utility functions. The focus in evaluating decision procedures is on value judgments (what is "desirable") and on empirical claims (about the likelihoods of various outcomes).

6.1.1. Standard Setting

Many test-based decision procedures involve one or more cutscores that define the decision rule (Cizek, 2001). For example, states use licensure tests to decide whether candidates for professional licensure should be allowed to engage in practice. In most cases, it is not feasible to evaluate performance in practice directly, and the standard approach is to assess professional competency over a target domain of cognitive skills that are considered critical for effective performance in practice (AERA et al., 1999). A single cutscore, or passing score, is defined on the score scale for the licensure test. Candidates who pass the licensure test and meet all other requirements get licensed; candidates who do not pass the test do not get licensed (Clauser, Margolis, & Case, this volume).

The interpretive argument for a licensure test typically involves a semantic interpretation of professional competence as a broadly defined trait variable and then a decision procedure that implements a policy about the level of competence required for admission to practice. The validation of such trait interpretations is discussed in Section 3. The target domain of competencies considered critical for effective performance in practice is based on analyses of patterns of practice (LaDuca, 1994; Raymond, 2001).

The choice of cutscore is the main issue in defining the decision rule, and the evaluation of the cutscore is the key issue in validating the decision rule. Standard-setting studies are designed to identify a reasonable cutscore and to provide backing for the choice of cutscore (Hambleton & Pitoniak, this volume). In applying any standard-setting method it is necessary for the participants (e.g., state officials, professional practitioners, members of the public) to develop a conception of the level of competence needed for the intended use (e.g., safe and effective performance in practice). This conception of minimal competence, the *performance standard*, can be based on standards of performance in the practice of the profession (Kane, 2002b). Judgmental standard setting methods (e.g., the Angoff method) are designed to translate the performance standard into a specific score on the score scale for the test. Candidates with scores above this cutscore have presumably met the performance standard, and candidates with scores below this cutscore have presumably not met the standard.

The performance standard provides a description of a minimal level of competence needed in practice. The corresponding cutscore provides an operational definition of the decision rule. An evaluation of the decision rule would address the appropriateness of the proposed performance standard (e.g., as the basis for withholding a license to practice a profession) and of the relationship between the cutscore and the performance standard.

Once the performance standard is defined, various kinds of empirical evidence can be used to evaluate how well the cutscore represents the performance standard (Hambleton & Pitoniak, this volume; Kane, 1994), but the performance standard itself is evaluated in terms of its anticipated consequences. The standard for licensure decisions is supposed to be high enough to provide adequate protection to the public, but not so high as to unduly restrict access to practice. Although licensure programs have social consequences (e.g., the availability of professional services in various areas, the representation of minorities and other subgroups in the professions), the standard setting activities for licensure tests tend to focus on the level of individual competence needed for effective performance in practice.

6.2. Test-Based Accountability Programs

The traditional view of testing programs as essentially non-interactive monitoring devices has been replaced by a recognition that testing programs often have a major impact on those assessed (Crooks, 1988; Frederiksen, 1984; Madaus, 1988; Moss, 1998b), and more recently, by a conception of some standardized tests as the engines of reform and accountability in education. Current test-based school accountability programs use standardized tests to hold schools accountable for student progress. The state develops content standards and develops tests based on these standards. Many of the content standards (particularly those involving more ambitious, extended performances) may be excluded from the test specifications, because they are not amenable to paper-and-pencil testing. The subset of the state standards eligible for inclusion in the tests along with the standardized conditions of observation define the universe of generalization for the state testing program.

The advocates of accountability programs see them as a way to raise academic standards by focusing attention on demanding content. It is assumed that the test-based accountability program will focus attention on the areas tested, and more generally, on the State Standards, and that this will be accomplished without serious loss of attention to other areas.

Under the "No Child Left Behind" (NCLB) Act (NCLB, 2002), student scores are transformed to general achievement levels, intended to reflect different levels of performance (e.g., below basic, basic, proficient, advanced). The achievement levels are defined by cutscores on the score scale for the test. All scores below the basic cutscore are considered below basic, scores between the basic cutscore and the proficient cutscore are considered basic, etc. The goal is to define statewide standards of performance and to encourage all students to reach some predefined level of achievement (e.g., the proficient level). The reduction of the test scores to three or four achievement levels involves a substantial loss of information, but as discussed below, the system is not designed to produce precise information.

The achievement-level scores for each grade are aggregated over students in each school to yield the percentage of students at each achievement level in each grade within the school (and within various subgroups). In general, accountability programs apply rewards or sanctions to the schools based on their students' performances on the tests. The NCLB program focuses on sanctions for failure, particularly failure to achieve certain increases in the percentages of students at or above the proficient level. If a school fails to meet the targets for improvement, it is subject to sanctions.

In adopting high-stakes tests as a way of enhancing educational quality in certain content areas, the focus is not on monitoring achievement, but on improving it. The test is being used to implement a policy, which will have consequences, some of which are likely to be positive and some negative. To evaluate a testing program as an instrument of policy, it is necessary to evaluate its consequences (Cronbach, 1982).

6.2.1. An Interpretive Argument for NCLB Accountability Programs

Interpretive arguments for NCLB accountability programs would involve an initial semantic interpretation of student performance in terms of individual achievement on the state standards (a trait attribute), a conversion of these scores to achievement levels (basic, proficient, advanced), and the computation of the percentages at each level in each grade for a school (and for subgroups), followed by a decision about the school.

The interpretation of individual student performance in terms of achievement of the state standards involves the scoring of student responses, generalization over the universe of generalization, and extrapolation to the trait associated with overall achievement on the state standards.

The first two inferences, scoring and generalization, can be evaluated using the methods described in Section 3. The third inference, extrapolation, is potentially more problematic for accountability tests than for lower-stakes applications. Because the sanctions or rewards associated with the program are based on test scores, they will focus attention on the universe of generalization. As a result, the schools may give less attention to other areas of the curriculum. In particular, they may give less attention to those parts of the state standards not included in the test. There is nothing in the logic of these programs to suggest that achievement in any area not covered by the tests will improve as a result of the testing program. At best, performance on the topics not included in the tests should remain about the same.

Evidence for the extrapolation inference could be generated by examining the relationship between scores on the accountability tests and measures of achievement in areas not covered by the test. The extrapolation inference can also be evaluated by examining the impact of the testing program on the courses taken by the students and on the functioning of these courses. Does the attention given to areas not covered by the tests (e.g., art, music) change? In subjects covered by the accountability tests, does the focus shift away from content areas not covered on the test in favor of test-preparation activities?

Defining performance standards and cutscores for achievement levels introduces an array of value judgments (how good must a performance be in order to be considered proficient), which are made state to state (Linn, 2005). Unlike the cutscores for licensure examinations, which can be based on existing standards of practice, the educational accountability programs are not tied to any particular performance arena and therefore have no established basis for the standards. For example, should the definition of the proficient level in twelfth grade mathematics reflect the level needed in everyday life, in a trade school, or in a college engineering program?

The introduction of value judgments is not necessarily a problem. Decision procedures always rest on value judgments. However, for validation, it is appropriate that the values be made explicit, and that the consequences of the decision rule be evaluated.

The shift from individual student scores to school-based percentages at or above the proficient level shifts the focus from individual students to schools. The school-level results will be skewed if some schools fail to test many low-scoring students. To address this issue, NCLB includes participation rules.

Finally, the NCLB accountability program imposes various sanctions on schools that fail to meet specified requirements for increases in the percentage of students at or above the proficient level. A distinguishing feature of test-based accountability programs is their focus on achieving certain goals, in addition to or instead of focusing on the measurement of any particular attributes. This is particularly true of the NCLB legislation. The provisions of this Act include mandates on when testing is to occur (grades 3 through 8), which students are tested (requirements on participation rates for various groups), and consequences for schools, but the act defers to state standards on the content and format of the tests and on the definition of the achievement levels (Linn, 2005). NCLB says that schools have to test their students and it specifies how the scores are used for accountability, but it leaves the semantic interpretation of the test results to the states.

Test-based accountability programs have a range of potential benefits and costs (Kane, 2002c; Lane, Parke, & Stone, 1998; Linn, 1993; Mehrens, 1997). The potential benefits include increases in student achievement on the content areas covered by the tests (Linn, 2005), increases in achievement on the State Standards, and, possibly, improvements in public confidence in the schools. The potential costs include the time and resources spent on testing, possible narrowing of the curriculum and of student options (e.g., fewer AP courses), and increased dropout rates. These positive and negative consequences are likely to have different impacts on different groups and in different schools (Lane & Stone, 2002).

It is possible to estimate the extent to which the intended outcomes of the testing program actually occur (e.g., higher levels of achievement on the State Standards). It is also possible to estimate the extent to which specific unintended outcomes (e.g., increased dropout rates, decreases in elective course offerings and enrollment, declines in participation rates in extracurricular activities) occur after

the introduction of the testing program (McNeil, 2005). Evidence relevant to these questions could also be derived from studies of changes in the schools after the tests are introduced.

The central question is the extent to which the accountability program has an impact (positive or negative) on achievement and other outcomes. To answer this question would require a program evaluation. The accountability program is functioning as an educational intervention and would be evaluated as such.

6.3. The Role of Consequences in Validity

Consequences have always been a part of our conception of validity (Guion, 1974; Messick, 1975, 1998; Shepard, 1997). Traditional definitions of validity in terms of how well a testing program achieves its goals (Cureton, 1951) necessarily raise questions about consequences, positive and negative (Cronbach & Gleser, 1965, Linn, 1997; Moss, 1992). Test-based decision procedures are necessarily evaluated in terms of their outcomes, or consequences, but until the 1970s, the consequences being evaluated tended to be mainly local, direct benefits and costs (e.g., increased efficiency vs. dollar costs).

The civil rights movement of the 1960s brought issues of fairness and equity to center stage (Cole & Moss, 1989; Ebel, 1966), and legislation and judicial decisions changed concerns about social equity into legal issues. Cronbach (1980a) described an environment in which every established testing practice was being criticized by someone who "disliked its consequences" (p. 37).

In 1978, the federal agencies responsible for enforcing civil rights legislation published the *Uniform Guidelines on Employee Selection Procedures* (EEOC et al., 1978). The Uniform Guidelines stated that if a selection procedure or a component of a selection procedure produced adverse impact against a protected group (e.g., racial minorities), the procedure could not be used for employment decisions unless the test scores were shown to be related to performance on the job. Adverse impact was said to exist if the selection rate for the protected group was less than four-fifths of that for the majority group.

Under the Uniform Guidelines, a finding of adverse impact triggers a requirement that the test user "validate" the test for the proposed use (e.g., by showing that test scores are empirically related to job performance). Essentially, adverse impact raises the stakes and creates a presumption of possible bias, which can be answered by supplying evidence that test scores are job related. The Uniform Guidelines struck a balance between the employer's interest in hiring the best candidates for each job and society's interest in group equity.

At about the same time, the courts decided that the state cannot deprive a student of a diploma on the basis of test scores without demonstrating that the student had an opportunity to master the competencies being measured (*Debra P. v. Turlington,* 1983; Linn, 1989). This decision can be viewed as a compromise between the state's interest in promoting high educational standards and students' rights under the Fourteenth Amendment (Jaeger, 1989). In these judicial decisions, the focus was on the balancing of positive and negative consequences.

Messick (1975, 1989, 1995) made social consequences a major issue in his analyses of validity (Moss, 1992). Cronbach (1971, 1980a, 1988) also gave a lot of attention to consequences in his discussions of validity but did not make them an organizing dimension in his analyses. In reaction to Messick's emphasis on the consequential aspect of validity, some authors (Borsboom, Mellenbergh, & van Heerden, 2004; Mehrens, 1997; Popham, 1997; Sackett, 1998) have argued for a more limited definition of validity, involving primarily the semantic interpretation of scores. For example, Popham (1997) acknowledged that consequences were important but preferred not to consider them part of validity. Linn (1997) and Shepard (1997) favored a broader conception of validity, which would include an evaluation of both the meaning of test scores and the consequences of their use.

6.3.1. The Role of Consequences in Evaluating Decision Procedures

Consequences, or outcomes, are the bottom line in evaluating decision procedures, which are always designed to achieve some desired outcomes or to avoid some undesirable outcomes (Cronbach, 1971; Shepard, 1993, 1997). A decision procedure that does not achieve its goals, or does so at too high a cost, is likely to be abandoned, even if it is based on perfectly accurate information. In medicine, a highly accurate diagnostic procedure for an untreatable illness would not be very useful as a screening tool, especially if the diagnostic procedure were costly, painful, or at all risky. In applied settings the bottom line involves a weighing of positive and negative consequences (Cronbach & Gleser, 1965; Debra P. v. Turlington, 1983; EEOC et al., 1978; Jaeger, 1989).

Evidence for the accuracy of the information is certainly relevant to the evaluation of a decision procedure but mainly because more accurate information is expected to lead to better decisions. If the information is shown to be erroneous, confidence in the decisions is likely to collapse. However, even indisputable evidence for accuracy does not justify a decision procedure (Cronbach, 1971, 1988).

As noted earlier, the evaluation of how well a decision procedure achieves its goals and of the immediate negative consequences or costs of testing have always been an integral part of the validation of decision procedures, but social consequences (particularly adverse impact) did not get much attention until the 1970s. Although there is general acceptance within the measurement community of the use of social consequences by the courts and other government agencies in evaluating test-based decision procedures, there has been some debate about whether the evaluation of social consequences should be included under the heading of validity.

Test-based accountability programs (e.g., "No Child Left Behind") blur the distinction between intended consequences and social consequences by adopting as their major purpose the improvement of educational outcomes for all students (Haertel, 1999). These testing programs have

moved beyond the traditional monitoring role, to the use of testing as the engine of reform and accountability in education. Since these testing programs are intended to improve (or "reform") educational institutions, it seems reasonable to evaluate them as educational programs. Program evaluations include the evaluation of intended and unintended outcomes of the program being evaluated.

6.3.2. Impact of Consequences on Attribute Definition

Conclusions about the consequences associated with a decision procedure can impact the evaluation of a measurement procedure in two ways. First, information about specific consequences can pinpoint weaknesses or problems in the measurement procedure (Messick, 1989). For example, a finding that many students are being placed into courses that are at too high a level because the placement test does not give enough attention to some critical skill would suggest that the test be modified to give more weight to that skill. Research on differential item functioning (DIF) essentially uses information on group differences in performance to identify items that may be flawed.

Second, information about consequences provides evidence about the appropriateness of the measurement procedure for the purpose at hand. If the decision procedure is not working as intended (e.g., many students are being placed into the wrong course), but no specific problems in the measurement procedure can be identified, it may be reasonable to conclude that the attribute that is being measured is inappropriate (Cronbach, 1971). For example, using a general academic aptitude test to place students into a sequence of engineering courses might not work very well, because the different courses require specific mathematical skills that are not being measured. There is nothing wrong with the test of academic aptitude, but the test does not provide the information that is most critical to effective placement decisions. It is the wrong test for this purpose.

6.3.3. Responsibility for Evaluation of the Consequences

Having drawn a distinction between the semantic interpretation and the decision rule, it is often possible to evaluate the two parts of the interpretive argument separately (Reckase, 1988). Test developers are the obvious candidates to validate the claims they make (explicitly or implicitly in labels and suggested uses) in the proposed semantic interpretation of test scores. Many of the semantic inferences are routinely evaluated in test development (e.g., evaluation of scoring rules, estimation of generalizability).

Test users (i.e., those who make the policy decision to use a test for some purpose) should play a large role in analyzing the consequences of a test use. Test users identify the kinds of decisions to be made and the procedures to be used to make these decisions (Cronbach, 1980b; Taleporos, 1998). They presumably know the intended outcomes, the procedures being employed, and the population being tested, and therefore, they are in the best position to identify the intended and unintended consequences that occur. The test users may choose to rely on evidence and analyses provided by a test developer as one source of information, and it would be reasonable to expect that test developers would provide support for any test uses that they recommended, explicitly or implicitly. (Shepard, 1997).

In spite of the conceptual distinction between the semantic interpretation and the decision and the possibility of having different groups be responsible for the two parts, it is the interpretive argument as a whole that is to be evaluated. Assuming that a test has been found to assess some trait and that a measure of that trait is to be used to make certain decisions, it is not necessarily the case that the procedure will be effective. Aspects of the procedure, the context, or the population being tested may interfere with the effectiveness of the procedure in a particular context. For example, in a high-stakes testing environment, instruction that focuses on the content and format of test tasks (i.e., "coaching") may lead to changes in the meaning of test scores. In particular, the scores may become less indicative of a broad content area to the extent that practice on test questions replaces more general instruction (Heubert & Hauser, 1999). Tests that are susceptible to such subversion may work very well in low-stakes contexts but fail in high-stakes contexts.

Each application of a measurement procedure has to be evaluated on its own merits. Many different kinds of evidence may be relevant to the evaluation of the consequences of a testing system (Lane, Parke, & Stone, 1998), and many individuals, groups, and organizations may be involved in generating the evidence (Linn, 1998).

6.3.4. Social Consequences

Cronbach and Messick have suggested very different approaches to the inclusion of social consequences within their models of validity (Moss, 1992). Messick described the appraisal of social consequences as an aspect of construct validity, because the outcomes of test use "both derive from and contribute to the meaning of test scores" (Messick, 1995, p. 7), and suggested that adverse social consequences invalidate test use mainly if they are due to flaws in the test:

> If the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced ... then the validity of the test use is not overturned. (Messick, 1989, p. 88)

It is generally agreed that adverse social consequences count against validity when they "can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components" (AERA et al., 1999, p. 16).

In a sense, to say that social consequences count against validity only when they are due to sources of invalidity is to give them a secondary role in validation. Under this model, adverse consequences serve mainly to suggest that some cases of invalidity are more serious than others. A threat to validity that might otherwise be ignored (e.g., one that has a small impact on individual test scores) becomes a serious concern if it is shown to have a systematic negative impact on some group. Legal analyses of test consequences and

validity have followed this pattern since the 1970s (EEOC et al., 1978). If a test has adverse impact on the hiring of minority workers, the validity of the test has to be demonstrated. If a satisfactory validation is produced, the adverse impact does not count against the legitimacy of the testing program.

For Cronbach (1971, 1988), consequences had a more direct role in validation. He suggested that negative consequences could invalidate test use even if the consequences cannot be traced to any flaw in the test, because "tests that impinge on the rights and life chances of individuals are inherently disputable" (Cronbach, 1988, p. 6). Cronbach (1988) argued that:

> Validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to argue against adverse consequences. (p. 6)

He concluded that we

> may prefer to exclude reflection on consequences from meanings of the word *validation*, but ... cannot deny the obligation. (p. 6)

Messick (1989) took exception to Cronbach's flexibility on this matter of definition. After quoting Cronbach's (1988) views, Messick (1989) argued that:

> the meaning of validation should not be considered a preference. On what can the legitimacy of the obligation to appraise social consequences be based if not on the only genuine imperative in testing, namely, validity? (p. 20)

Cronbach (1988) was less insistent than Messick (1989, 1995) about including the evaluation of consequences under validation but was inclined to consider all consequences in evaluating the legitimacy of test use.

Validation has a contingent character in that the evidence required for validation depends on the proposed interpretations and uses. The test user has an obligation to make a case for the appropriateness of the decision procedure in the context in which it is being used. As is the case for all interpretive arguments, there is a ceteris paribus assumption inherent in test uses. In education, as in medicine, there is an obligation to avoid doing harm if it can be avoided. Therefore, the test user has an obligation to consider any negative consequences that can reasonably be anticipated and to weigh them against the potential benefits before adopting the test.

The critic has a right to challenge any assumption or inference in the interpretive argument, including those supporting semantic inferences and those supporting claims about the benefits to be expected from a proposed test use, and the test users should be prepared to defend their choices. All test users have to operate within the bounds of the law, and most test users are also constrained by public opinion. They cannot ignore consequences, even if they would prefer to do so. On the other hand, it is not possible to anticipate all possible consequences of any decision, so it would be counterproductive to demand that all consequences be considered before a decision procedure can be implemented.

Toulmin (1958) has provided a reasonable resolution to this dilemma. He suggested that the general rules of argumentation apply over a wide range of contexts, but that the criteria for evaluating real arguments tend to depend on the context in which they occur. The kinds of consequences to be considered in evaluating a decision procedure vary as a function of the purposes of the procedure and the context in which it operates. So for example, in making selection decisions, businesses tend to be concerned about gains in productivity (a positive consequence) and about costs (a negative consequence). In making placement decisions, schools are concerned about optimizing learning vs. the negative consequences of failure. Licensure organizations seek to protect the public, while not excluding qualified candidates.

Any consequences that are considered significant by stakeholders are potentially relevant to the evaluation of how well a decision procedure is working. The range of consequences that can be considered fair game has evolved over time. Before 1960, adverse impact was not given much attention; by 1980, adverse impact and racial, ethnic, and gender bias were clearly on the agenda. More recently the impact of standardized testing procedures on individuals with disabilities has become a major concern. The measurement community does not control the agenda; the larger community decides on the questions to be asked (Cronbach, 1988).

Cronbach (1988) has pointed out that, for validity arguments to be convincing to diverse audiences, the assumptions in these arguments must be credible to those audiences. This concern is especially salient for the policy assumptions implicit in high-stakes testing programs, because the credibility of these policy assumptions may be highly variable across stakeholders.

The evaluation of the policy assumptions inherent in high-stakes testing programs raises some difficult issues (Sackett, 1998), but they are not any more difficult than the issues routinely faced by program evaluators. To the extent that the purpose of a testing program is to promote certain outcomes, the testing program is functioning as an educational intervention and therefore merits an evaluation of the kind routinely mandated for new educational programs. For stakeholders to make informed decisions about the effectiveness of high-stakes tests, it is necessary that they have information about how well these tests achieve various goals and at what cost. Assuming that there are both positive and negative consequences, the stakeholders and policymakers face the task of weighing these consequences against each other.

7. FALLACIES IN VALIDITY ARGUMENTS

Validity arguments can go wrong in many ways. Any empirical study included in the validity argument can be designed or carried out poorly, and any empirical results can be misinterpreted or misapplied. On a larger scale, the interpretive argument can be misspecified in the sense that it does not correspond to the patterns of actual score interpretations and uses. These misspecifications of the interpretive argument tend to be more serious than technical

problems in a particular study, because they are more pervasive and harder to detect.

A fallacy occurs when an unsound argument appears to be sound. Faulty reasoning may appear sound on casual inspection, and if it is subtle enough, it can mislead both proponent and audience (Hansen & Pinto, 1995). It is one of the tasks of validation research to expose any gaps in the formulation of interpretive arguments.

There are at least three ways in which interpretive arguments can be misspecified, each of which is a special case of a classic fallacy: begging-the-question, the straw-man fallacy, and gilding the lily.

7.1. Begging the Question

A common problem in validating interpretive arguments is the tendency to take one or more questionable inferences for granted or to take part of the interpretive argument to be the whole of the interpretive argument. This fallacy has traditionally been referred to as "begging the question," because the question at issue, or a large part of the question at issue, is simply taken for granted (Walton, 1989).

Many real-world applications of testing assume fairly ambitious interpretations of test scores, involving many inferences and assumptions. That some of these inferences and assumptions have not been systematically evaluated at a given point in time is not necessarily a problem, especially if they are reasonably plausible. However, it is a problem if the unexamined assumptions and inferences are not plausible, and it is especially dangerous if the implausible inferences are implicit and rest on hidden assumptions.

As noted in Section 1.2.1, traditional analyses of "content validity" tended to beg questions about implicit trait or construct interpretations. Claims for content validity were often based on studies that evaluated only one assumption, the relevance of items to the attribute being measured, while a sound interpretive argument for an observable attribute generally requires at least three inferences: scoring, generalization, and extrapolation. The traditional content-validity analysis could provide support for the scoring inference. In addition, expert opinion indicating that the sample of performance is representative of the universe of generalization combined with empirical evidence for generalizability over this universe can provide adequate support for generalization to the universe score. However, expert evaluations of test items do not generally provide strong support for extrapolation to the target domain, while the interpretations and uses of the scores typically assume such extrapolation.

The begging-the-question fallacy fits with the confirmationist tendency in many validation studies. For example, the proponents of authentic assessments have tended to emphasize the extent to which the performances observed in testing match the target performance while taking the generalization of observed scores over tasks, occasions, and conditions of observation for granted, even though empirical research consistently indicates that generalizability over performance tasks cannot be taken for granted. Similarly, developers of objective tests have tended to give a lot of attention to content representativeness and generalizability over items, while taking extrapolation to the target performance for granted. In both cases, the more questionable part of the argument is accepted without much critical scrutiny. The fact that some major inferences are being glossed over may not be noticed, especially if other parts of the argument are developed in some detail.

7.1.1. Begging the Question of Consequences

This fallacy is a special case of the begging-the-question fallacy, but it occurs frequently enough that it deserves separate discussion. It is generally inappropriate to assume that evidence supporting a particular interpretation of test scores automatically justifies a proposed use of the scores. That test scores provide accurate estimates of a relevant student attribute does not generally imply that they will be useful to the student's teacher in planning instruction (Tittle, 1989).

A classical example of this fallacy was provided by Cronbach and Snow (1977) in their analysis of aptitude-treatment interactions. Assume, for example, that a test is an excellent predictor of performance in two treatment options, A and B, but that the regression lines for outcomes based on scores are parallel, so that everyone does uniformly better in treatment A than in treatment B. In this case, the test scores are not in themselves at all useful for placement decisions; the optimal policy is to assign everyone to treatment A. The validity of the interpretation as a predictor of performance in the two treatment options does not support the validity of the proposed placement decisions.

Shepard (1993) provides another good example. Suppose that a "readiness" test is an excellent predictor of performance in kindergarten. Does this justify its use in deciding whether to admit children to kindergarten this year or hold them back till next year? If a low "readiness" score indicates a developmental lag that will be resolved by waiting a year, this strategy would make sense. If the low score indicates a home environment that does not promote learning of the skills, keeping the child out of school for another year would be counterproductive. In any case, the validity of the proposed interpretation in terms of mastery of certain skills does not justify the use of the test scores for what is essentially a placement decision (kindergarten or home).

Current accountability programs provide many examples of begging the question. The arguments for these testing programs tend to claim that the program will lead to improvements in school effectiveness and student achievement by focusing the attention of school administrators, teachers, and students on demanding content. Yet, the validity arguments developed to support these ambitious claims typically attend only to the descriptive part of the interpretive argument (and often to only a part of that). The validity evidence that is provided tends to focus on scoring and generalization to the content domain for the test. The claim that the imposition of the accountability requirements will improve the overall performance of schools and students is taken for granted.

The evaluation of the policy assumptions inherent in accountability programs and other high-stakes testing programs requires an evaluation of intended and unintended outcomes. If the primary purpose of a testing program is to promote certain outcomes (e.g., achievement of the

demanding content outlined in the state standards) rather than to measure certain variables, the assumption that the testing program will yield these outcomes deserves some attention.

Scriven (1995) has made a similar point in the context of program evaluation. He suggested that a well designed and implemented evaluation effort can lead to well-founded conclusions about a product or program, but that in general, it is not appropriate for the evaluator to make recommendations based on these conclusions. Even if the evaluative conclusions are rock solid given the data, the actions to be taken by decision makers generally depend on many considerations not addressed in the evaluation. In particular, unless the evaluators have examined the potential consequences of different actions that might be taken and the values of the relevant stakeholders, they are not in a position to identify the best course of action.

7.1.2. Overgeneralization or Spurious Precision

The social sciences have, at times, been led astray by trying to mimic the physical sciences too closely, but there is at least one area in which it might be beneficial to follow the example of the physical sciences. Physical scientists are expected to investigate all sources of error that might have a substantial impact on their measurements. As Cronbach et al. (1972) pointed out, reliability studies often implicitly define the universe of generalization too narrowly given their measurement procedures, and in doing so, "they underestimate the 'error' of measurement, that is, the error of generalization" (p. 352).

Generalizability analyses can be used to examine the magnitudes of different sources or random sampling error in any measurement procedure, including the errors due to the sampling from the task, occasion, context, administrator, or rater facets (Brennan, 2001b; Cronbach et al., 1972; Haertel, this volume). Questions about precision are begged if the random errors associated with generalization over some facets are taken to represent unbiased estimates of the total error in generalizing to the expected value over the universe of generalization. In such cases, the estimated random errors underestimate the sampling errors for the full set of facets in the universe of generalization. For example, the use of data from a single administration of an objective test to estimate a standard error based on coefficient alpha includes information on the sampling error for items but does not provide any indication of the impact of other potential sources of error. This limited estimate of sampling error is likely to be an underestimate of the total error in generalizing over occasions, raters, contexts, and other facets.

It may be reasonable to simply assume invariance over some facets that are not expected to have much influence on performance (e.g., the room used for a multiple choice test), and these facets do not necessarily have to be investigated. In addition many facets can be included in the residual error by allowing them to vary randomly in the G study. However, those facets that are expected to introduce substantial error into the generalization inference deserve detailed analysis, and to the extent that some potentially serious sources of error (random or systematic) have not been examined, these limitations should be acknowledged.

7.1.3. Surrogation

Just as correlation does not imply causality, even a very high correlation between two measures does not imply that they have the same meaning. Scriven (1987) has labeled the "use of a correlate ... as if it were an explanation of, or a substitute for, or a valid evaluative criterion of, another variable" as the *fallacy of statistical surrogation* (p. 11). The fallacy involves a "substitution of a statistical notion for a concept of a more sophisticated kind such as causation or identity" (Scriven, 1987, p. 11). Among students who have graduated from American high schools, the correlation between scores on a mathematics test and an English usage test would generally be positive and fairly high; but level of achievement in English is not the same thing as level of achievement in mathematics. The high correlation between English and math scores reflects the fact that successful students typically do well on both tests, and failing students do less well on both tests. However, if math were dropped from the standard high-school curriculum or students from all over the world were included in the analyses, these correlations would probably drop sharply.

Scriven introduces the surrogation fallacy in the context of employee evaluation and distinguishes "primary indicators," which involve samples of the performance of interest from "secondary indicators," which are "only statistically connected with good (or bad) job performance" (Scriven, 1987, p. 17). He points out that it would be inappropriate to dismiss someone from a job because of rumors that they are an alcoholic, even if a statistical relationship can be shown between rumors of this kind and alcoholism, and between alcoholism and poor performance.

The surrogation fallacy is particularly dangerous in high-stakes contexts, because in such contexts, it is predictable that, if the indicators can be manipulated, they will be manipulated. If certain indicators are used for selection decisions because they have been found to be good indicators of future performance, they may not be good indicators for long. If scores on a vocabulary test are used as indicators of language skills in college admissions, applicants will practice this skill, and its value as an indicator of overall verbal competence (e.g., reading, writing, speaking, listening) is likely to decline. Large-scale testing programs tend to be highly standardized to promote fairness and tend to rely on objective formats for practical reasons. Under these circumstances, specific preparation for test tasks is likely to increase and preparation for other tasks is likely to decrease. Test preparation will focus on the formats included in the test in preference to activities not included in the test. The higher the stakes, the more likely it is that the surrogate measures will displace the performances of interest.

7.1.4. Reification

Reification involves a leap from an observed regularity in scores to the existence of some "thing" that is the source of the regularity. For example, it is natural to assume that observed consistencies in performance over some domain of tasks correspond to (and even are caused by) some trait that exists in persons. This tends to be especially true in cases where the performances in the domain all share some

structural characteristics (e.g., series completion tasks), and we have some general ideas about how people perform such tasks (e.g., by identifying the pattern in the series and completing it). In such cases, it is natural to talk about "analytic ability," and there is no great harm in doing so as long as we refrain from imbuing the term with a lot of excess meaning (e.g., that it is hereditary, or that it is a prerequisite for success in science). As discussed in Section 3, the scores derived from standardized assessments are interpreted in terms of expected performance over the universe of generalization, and by extrapolation, in terms of the target score. Similarly, IRT models yield estimates of latent abilities derived from performance in some domain of items. Assuming that the data fit the model, the model justifies an interpretation of scores in terms of level of performance in the domain. However, causal inferences about an underlying trait that accounts for the observed regularities requires additional evidence.

Factor analysis provides a framework for analyzing relationships among different domains in terms of factor loadings. The factor loadings summarize the observed pattern of correlations among scores but do not necessarily correspond to any fundamental set of traits. The patterns of factor loadings may in fact be due to common patterns in the sequencing of instruction or other characteristics of the instructional or testing environment.

The point here is not to argue against the use of trait language. Thinking about traits or constructs as underlying causal variables can be a fertile source of insight about behavior and can yield hypotheses for further analysis, and various kinds of statistical analyses can provide useful guidance in defining traits and in defining their target domains. However, if excess meaning is to be brought into an analysis in the form of hypothesized traits or constructs, the claims being made, explicitly or implicitly, should be examined.

7.2. The Straw-Man

The begging-the-question fallacy takes some inferences and assumptions for granted. The second fallacy, the "straw-man fallacy," goes in the opposite direction and adopts an interpretive argument that is more ambitious than that required for the proposed interpretations and uses.

Given the potential for unnecessary difficulties in overstating the interpretation and thereby making it more ambitious than it needs to be, why do interpretations get overstated? I suggest three possible reasons.

First, the details of the interpretive argument implicit in a proposed test interpretation or use are not necessarily obvious in most cases. For example, licensure examinations can be interpreted as measures of current competence or as predictors of future performance. If one does not consider the various options carefully, it is easy to run them together and employ the stronger, predictive model when the competency model is sufficient. The proposed interpretations and uses can easily be overstated through carelessness.

Second, there is some tendency for test developers and vendors to propose ambitious interpretations. Test development projects often start with lofty goals, and correspondingly ambitious interpretive arguments. As development proceeds, compromises may become necessary and some parts of the plan may not get completed, but footnotes and caveats tend to get lost in general statements of the proposed interpretations and uses. In the case of legally mandated testing (federal or state sponsored), there is a tendency for political rhetoric to greatly inflate the claimed benefits of testing programs; if these claims were taken seriously, they would imply the need for a very ambitious validation effort.

Third, individuals who are inclined to challenge a proposed interpretation or use have a natural tendency to state the underlying interpretation of scores in a way that is easy to refute, and overly ambitious assumptions are easy to knock down. So it is tempting for hostile critics to treat the proposed interpretation or use as a "straw man" by overstating its assumptions.

7.3. Gilding the Lily

Any argument can be made to appear more convincing without being substantially strengthened by adding additional evidence for inferences and assumptions that are already quite plausible. This fallacy, referred to as "gilding the lily," is not particularly harmful in itself but can be pernicious when it occurs in connection with other fallacies. It is easier to ignore the fact that support for some key inference is weak, if an extensive array of evidence is provided for other, more plausible inferences or assumptions.

For example, in validity analyses for multiple-choice test scores, it is not unusual to find multiple analyses of generalizability over items, but little or no evidence for more questionable inferences (e.g., extrapolations to performance domains or traits). For objective testing programs, it is generally quite easy to generate internal-homogeneity reliability coefficients (e.g., coefficient alpha), and it is tempting to accumulate lots of this easily available and reassuring evidence. And there is nothing wrong with this in itself.

The problem arises if the accumulation of evidence in support of a relatively safe assumption distracts test developers and users from critically examining other less plausible inferences. For a well developed multiple-choice test with a substantial number of items, it would be surprising if coefficient alpha were not fairly high. Conducting a study to confirm generalizability over items is certainly in order, and monitoring it over time is also appropriate, but it should be recognized that multiple studies of this kind do not add much to our confidence in the interpretive argument as a whole.

8. CONCLUDING REMARKS

Validation involves the evaluation of the proposed interpretations and uses of measurements. The interpretive argument provides an explicit statement of the inferences and assumptions inherent in the proposed interpretations and uses. The validity argument provides an evaluation of the coherence of the interpretive argument and of the plausibility of its inferences and assumptions. It is not the test that is validated and it is not the test scores that are validated. It

is the claims and decisions based on the test results that are validated. Therefore, for validation to go forward, it is necessary that the proposed interpretations and uses be clearly stated.

Over the last fifty years, three general principles of validation have developed out of construct validity. The proposed interpretation is to be stated clearly enough to indicate what it implies. The evaluation of the interpretive argument is to involve an extended analysis of the proposed interpretations and uses. Plausible alternative interpretations are to be identified and investigated. The mix of evidence to be used in evaluating a particular interpretive argument depends on the claims being made by that interpretive argument.

For high-stakes testing programs, agreement on the interpretations and uses may require negotiations among stakeholders about the conclusions to be drawn and the decisions to be made (Ryan, 2002). The interpretation of a measure of competence in some area (e.g., writing), may seem simple enough, but in defining the competence and in developing the test, a number of specific issues typically arise, including the kinds of tasks to be included (the writing of short essays on demand or the production of longer works over an extended period), the scoring rules (relative emphasis on the mechanics, organization, style, and creativity), and the conditions under which the performance will occur (e.g., with or without tight time constraints).

The argument-based model provides a relatively pragmatic approach to validation. The goal is to develop a measurement procedure that supports the proposed interpretations and uses and an interpretive argument that is plausible, given the measurement procedure. The initial development effort is legitimately confirmationist. During the appraisal stage, the proposed interpretations and uses are taken as hypotheses or conjectures that are to be subjected to critical evaluation.

An explicitly stated interpretive argument serves three critical functions. First, it provides a framework for test development by indicating the assumptions that need to be met. The development of a measure of current competence in a content area would be different in many ways from the development of an indicator of a theoretical construct.

Second, the interpretive argument provides a framework for the validity argument. The evidence called for in the validity argument is that needed to support the specific inferences and assumptions in the interpretive argument.

Third, the interpretive argument provides a basis for evaluating the adequacy of the validity argument. If the validity argument provides adequate support for the interpretive argument, the proposed interpretations and uses are warranted. If the interpretive argument is incomplete or some of its inferences or assumptions are shaky, some interpretations or uses are not fully justified.

Validation has a contingent character; the evidence required to justify a proposed interpretation or use depends on the proposed interpretation or use. A modest interpretation (e.g., in terms of an observable attribute) makes fewer claims and is therefore easier to justify than a more ambitious interpretation (e.g., in terms of a broadly defined trait or construct).

The fact that we develop the interpretation, rather than discover it, has two major implications. First, we have some flexibility in what we choose to put into the interpretation. Second, the fact that the interpretation is not a given implies that it needs to be specified clearly if it is to be evaluated effectively. The specification of an interpretive argument puts a definite proposal on the table. The proponent has a claim to defend, and the critic has one to challenge. The kinds of evidence required for the validation of a proposed interpretation or use depends on the structure of the corresponding interpretive argument.

In developing a validity argument, the proponent can make a positive case for a proposed interpretation by stating the interpretive argument clearly, thereby demonstrating its coherence, and by providing support for its inferences and assumptions. The critic can challenge the appropriateness of the proposed interpretations and uses of the test results, the adequacy of the interpretive argument given the goals of the testing program, or the plausibility of specific inferences or assumptions. The critic can propose an alternative assessment procedure or interpretation. A critic might also claim that a test-based decision procedure fails to produce the intended outcomes or has serious unintended negative consequences.

A failure to state the proposed interpretations and uses clearly and in some detail makes a fully adequate validation essentially impossible, because implicit inferences and assumptions cannot be critically evaluated. Most fallacies in presumptive reasoning involve the tacit acceptance of doubtful assumptions, and an important function of external critics and alternative interpretations is to make these assumptions explicit.

NOTE

1. The model being proposed here is essentially the same as that in Kane (1982), but the terminology has been changed to make it more natural and more consistent with standard usage in generalizability theory. In the terminology used here, the broader set of observations about which conclusions are to be drawn is called the "target domain," and the narrower set defining the standardized measurement procedure (the set from which observations are drawn and to which statistical generalization is legitimate) is called the "universe of generalization." These two sets approximate what I called the "universe of generalization" and the "universe of allowable observations" respectively in Kane (1982).

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin Supplement, 51*(2) 1–38.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37,* 1–15.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 9–13). Hillsdale, NJ: Lawrence Erlbaum.

Bachman, L. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21*(3), 5–18.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17*(1), 10–22.

Benson, J., & Hagtvet, K. (1996). The interplay between design, data analysis and theory in the measurement of coping. In M. Zeidner & N. Endler (Eds.), *Handbook of coping: Theory, research, applications* (pp. 83–106). New York: Wiley.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–73.

Blair, J. (1995). Informal logic and reasoning in evaluation. *New Directions for Evaluation: Reasoning in evaluation: Inferential links and leaps, 68,* 71–80.

Bonner, S. (2005, April). *Investigating the cognitive processes in responding to MBE questions.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Brennan, R. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*(4), 285–317.

Brennan, R. (2001b). *Generalizability theory.* New York: Springer-Verlag.

Bridgeman, P. (1927). *The logic of modern physics.* New York: Macmillan.

Brookhart, S. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5–12.

Bruner, J. (1990). *Acts of meaning.* Cambridge, MA: Harvard University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Carroll, J. (1986). What is intelligence? In R. Sternberg & D. Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definition* (pp. 51–54). Norwood, NJ: Ablex Publishing.

Cizek, G. (2001). *Standard setting: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum.

Clauser, B. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24,* 310–324.

Clauser, B. E., Harik, P., & Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement, 37,* 245–262.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). New York: American Council on Education and Macmillan.

Cook, T., & Campbell, D. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

Cronbach, L. J. (1980a). Selection theory for a political world. *Public Personnel Management, 9*(1), 37–50.

Cronbach, L. J. (1980b). Validity on parole: How can we go straight? *New Directions for Testing and Measurement: Measuring Achievement Over a Decade, 5,* 99–108.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

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

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods; A handbook for research on interactions.* New York: Irvington Publishers.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438–481.

Crooks, T., Kane, M., & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education, 3,* 265–285.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Debra P. v. Turlington, 644 F. 2d 397, 5th Cir. 1981: 564 F. Supp. 177 (M.D. Fla. 1983).

Delandshere, G. (2002). Assessment as inquiry. *Teacher's College Record, 104,* 1461–1484.

Downing, S., & Haladyna, T. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38,* 327–333.

Dwyer, C. A., Gallagher, A., Levin, J., & Morley, M. E. (2003). *What is quantitative reasoning? Defining the construct for assessment purposes* (Research Report 03–30). Princeton, NJ: Educational Testing Service.

Ebel, R. (1961). Must all tests be valid? *American Psychologist, 16,* 640–647.

Ebel, R. (1966). The social consequences of educational testing. In A. Anastasi (Ed.), *Testing problems in perspective: Twenty-fifth anniversary volume of topical readings from the Invitational Conference in Testing Problems* (pp. 18–29). Washington, DC: American Council on Education.

Eisner, E. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice.* New York: Macmillan.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. (1984). A general multicomponent latent trait model for measuring learning and change. *Psychometrika, 49,* 175–186.

Embretson, S. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380–396.

Embretson, S., & McCollam, K. (2000). Psychometric approaches to understanding and measuring intelligence. In R. Sternberg (Ed.), *Handbook of Intelligence* (pp. 423–444). Cambridge: Cambridge University Press.

Equal Employment Opportunity Commission (EEOC), Civil Service Commission, Department of Labor, & Department of Justice. (1978). Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures. *Federal Register, 43,* 38290–38315.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 237–270). Washington, DC: American Council on Education.

Flockton, L., & Crooks, T. (2002). *Social studies assessment results 2001.* Educational Assessment Research Unit. Dunedin, New Zealand: University of Otago.

Fournier, D. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions in Evaluation: Reasoning in Evaluation: Inferential Links and Leaps, 68,* 15–32.

Frederiksen, J. (2003). Issues for the design of educational assessment systems. *Measurement: Interdisciplinary Research and Perspectives, 1,* 69–73.

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

Galison, P. (1987). *How do experiments end?* Chicago: University of Chicago Press.

Geertz, C. (1973). *The interpretation of cultures.* New York: Basic Books.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests.* New York: Wiley.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24,* 355–392.

Grandy, R. E. (1992). Theory of theories; a view from cognitive science. In J. Earman (Ed.), *Inference, explanation, and other frustrations: Essay in the philosophy of science* (pp. 216–233). Berkeley: University of California Press.

Greeno, J., Pearson, P., & Schoenfeld, A. (1997). Implications for the National Assessment of Educational Progress of research on learning and cognition. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Assessment in transition: Monitoring the nation's educational progress. Background studies* (pp. 151–314). Stanford, CA: National Academy of Education.

Guion, R. (1974). Open a new window: Validities and values in psychological measurement. *American Psychologist, 29,* 287–296.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1,* 1–10.

Guion, R. M. (1980). On trinitarian conceptions of validity. *Professional Psychology, 11,* 385–398.

Guion, R. M. (1995). Comments on values and standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 25–27.

Guion, R. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum.

Gulliksen, H. (1950a). *Theory of mental tests.* New York: Wiley.

Gulliksen, H. (1950b). Intrinsic validity. *American Psychologist, 5,* 511–517.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Haertel, E., & Greeno, J. (2003). A situative perspective: Broadening the foundations of assessment. *Measurement: Interdisciplinary Research and Perspectives, 1,* 154–162.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York: American Council on Education and Macmillan.

Hansen, E., Mislevy, R., & Steinberg, L. (2003, April). *Evidence-centered assessment design and individuals with disabilities.* Paper presented at NCME Meeting, Chicago.

Hansen, H., & Pinto, R. (1995). *Fallacies, classical and contemporary readings.* University Park: Pennsylvania State University Press.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 130–159). Washington, DC: American Council on Education.

Heubert, J. P., & Hauser, M. H. (1999). *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: Nation Academy Press.

House, E. R. (1980). *Evaluating with validity.* Beverly Hills, CA: Sage Publications.

Ippel, M. J. (1986). *Component-testing: A theory of cognitive aptitude measurement.* Amsterdam: Free University Press.

Jaeger, R. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–534). New York: American Council on Education and Macmillan.

Jöreskog, K. (1973). A general method for investigating a linear structural equation system. In A. Goldberger & D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Academic Press.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6,* 125–160.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527–535.

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions, 17,* 133–159.

Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education, 9*(4), 355–379.

Kane, M. T. (2002a). Inferences about G-study variance components in the absence of random sampling. *Journal of Educational Measurement, 39*(2), 165–181.

Kane, M. T. (2002b). Practice-based standard setting. *The Bar Examiner, 71*(3), 14–24.

Kane, M. T. (2002c). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Klein, S., McCaffrey, D., Stecher, B., & Koretz, D. (1995). The reliability of mathematics portfolio scores: Lessons from the Vermont experience. *Applied Measurement in Education, 6*(1), 83–102.

LaDuca, A. (1994). Validation of professional licensure examinations. *Evaluation and the Health Professions, 17*(2), 178–197.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp 91–196). London: Cambridge University Press.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41,* 1183–1192.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23*(3), 6–14.

Lane, S., Parke, C., & Stone, C. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 24–28.

Lane, S., & Stone, C. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 23–30.

Linn, R. L. (1989). Current perspectives and future directions. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 1–10). New York: American Council on Education and Macmillan.

Linn, R. L. (1993). Linking results in distinct assessments. *Applied Measurement in Education, 6*(1), 83–102.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14–16.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 28–30.

Linn, R. L. (2005, June 28). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives, 13*(33). Retrieved June 28, 2005, from http://epaa.asu.edu/epaa/v13n33/

Linn, R., Kiplinger, V., Chapman, C., & LeMahieu, P. (1992). Cross-state comparability of judgments of student writing: Results from the New Standards Project. *Applied Measurement in Education, 5*(2), 89–110.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement, 3,* 635–694.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Madaus, G. F. (1988). The influences of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum* (pp. 83–121). *Eighty-seventh yearbook of the National Society for the Study of Education, Part I*. Chicago: University of Chicago Press.

McDonald, R. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum.

McDonald, R. (1999). *Test theory*. Mahwah, NJ: Lawrence Erlbaum.

McNeil, L. (2005). Faking equity: High-stakes testing and the education of Latino youth. In A. Valenzuela (Ed.), *Leaving children behind: How "Texas-style" accountability fails Latino youth* (pp. 57–111). New York: State University of New York Press.

Meehl, P. (1950). On the circularity of the law of effect. *Psychological Bulletin, 47,* 52–75.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16–18.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10,* 9–20.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103.). New York: American Council on Education and Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13–23.

Messick, S. (1995). Standards of validity and the validity of and standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research, 45,* 35–44.

Mislevy, R. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379–416.

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–62.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229–258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23,* 5–12.

Moss, P. A. (1998a). Recovering a dialectic view of rationality. *Social Indicators Research, 45,* 55–67.

Moss, P. A. (1998b). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12.

Moss, P. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22*(4), 13–25.

Moss, P. A., & Schutz, A. (2001). Educational standards, assessment, and the search for consensus. *American Educational Research Journal, 38*(1), 37–70.

Murphy, K. (2003). *Validity generalization: A critical review*. Mahwah, NJ: Lawrence Erlbaum.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1435 (2002).

Nystrand, M., Cohen, A., & Dowling, N. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment, 1*(1), 53–70.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "Two Disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research on Education, 24,* 307–353.

Pinto, R. (2001). *Argument, inference and dialectic*. Dordrecht, the Netherlands: Kluwer Academic.

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.

Popper, K. R. (1962). *Conjecture and refutation: The growth of scientific knowledge*. New York: Basic Books.

Porter, T. (2003). Measurement, objectivity, and trust. *Measurement: Interdisciplinary Research and Perspectives, 1,* 241–255.

Quine, W. (1953). Two dogmas of empiricism. In W. V. O. Quine (Ed.), *From a logical point of view* (pp. 20–46). New York: Harper and Row.

Raymond, M. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14,* 369–415.

Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13–16.

Rulon, P. (1946). On the validity of educational tests. *Harvard Educational Review, 16*(4), 290–296.

Ryan, K. (2002). Assessment validation in the context of high-stakes testing assessment. *Educational Measurement: Issues and Practice, 21*(1), 7–15.

Sackett, P. R. (1998). Performance assessment in education and professional certification: Lessons for personnel selection? In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating*

alternatives for traditional testing for selection (pp. 113–129). Mahwah, NJ: Lawrence Erlbaum.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education, 1,* 9–23.

Scriven, M. (1995). The logic of evaluation and evaluation practice. *New Directions in Evaluation: Reasoning in Evaluation: Inferential Links and Leaps, 68,* 49–70.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 19, pp. 405–450). Washington, DC: American Educational Research Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8, 13, 24.

Shepard, L. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 1066–1101). Washington, DC: American Educational Research Association.

Sireci, S., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75,* 457–490.

Snow, R. E., & Lohman, D. E. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–332). New York: American Council on Education and Macmillan.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.

Sternberg, R. J. (1979). The nature of mental abilities. *American Psychologist, 43*(3), 214–230.

Stiggins, R. J. (2005). *Student-involved assessment for learning* (4th ed.). Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Suppe, P. (1977). *The structure of scientific theories.* Urbana: University of Illinois Press.

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement: Issues and Practice, 17*(2), 20–23, 34.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum.

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

Tittle, C. (1989). Validity: Whose construction is it in the teaching and learning context? *Educational Measurement: Issues and Practice, 8*(1), 5–19, 34.

Toulmin, S. (1953). *The philosophy of science.* London: Hutchinson's Universal Library.

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Toulmin, S. (2001). *Return to reason.* Cambridge, MA: Harvard University Press.

van der Linden, W. (1998). A decision theory model for course placement. *Journal of Educational and Behavioral Statistics, 23*(1), 18–34.

Walton, D. (1989). *Informal logic: A handbook for critical argumentation.* Cambridge: Cambridge University Press.