Themes and Variations in Validity Theory

Pamela A. Moss
University of Michigan

Should the Standards *reflect the perspective that construct validity is central to all validation efforts? Is the construct-/content-/criterion-related categorization of validity evidence now obsolete? Should the definition of validity include consideration of the consequences of test use?*

As most of the readers of this journal know, the 1985 *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) is under revision. Deciding whether and how to revise the characterization of validity is probably the most significant issue facing the authors of the revised *Standards*. There is a substantial disjunction between the way validity is characterized in the 1985 *Standards* and in the published work of many who write about the philosophy of validity. While there are dominant themes that run throughout characterizations of validity in the *Standards* and elsewhere, there are also substantial variations, both with respect to the boundaries of validity—how it is delimited from other concepts—and with respect to its components—how it is analyzed into aspects, constituent parts, or processes to guide validity research.

The purpose of this article is to highlight some of the major questions facing the measurement community in deciding how to conceptualize validity in the revised *Standards* and thereby to encourage dialogue about how these issues might be resolved. Choices made with respect to the boundaries and components of validity are not just topics for the seminar room—they influence the way validity research is carried out, the responsibilities of assessment developers and users, and the rights of those who are assessed. This, in turn, can influence the kinds of assessments that are likely to find favor, the cost of developing, using, and evaluating assessments, and the impact of those assessments on various stakeholders, including the potential differential impact associated with concerns about fairness and equity.

The authors of the 1985 *Standards* took as one of their guidelines that "the *Standards* should . . . reflect the current level of consensus of recognized experts" (p. v.). In characterizing the themes and variations in the perspectives of recognized experts, I've chosen to focus primarily on the work of scholars who have written about the philosophy of validity in the context of educational measurement, although many of the issues I raise have been addressed by scholars writing primarily in other measurement contexts (e.g., Anastasi, 1986; Landy, 1986). In addition to the 1985 *Standards,* the pieces I draw on most heavily include the work of: Cronbach (1988, 1989), Haertel (1991, 1992), Linn (1993; Linn, Baker, & Dunbar, 1991), Kane (1992), Messick (1989a, 1989b, 1992, 1994a, 1994b), Shepard (1993), Wiley (1991; and Wiley & Haertel, in press). Clearly, the contributions of Cronbach and Messick are seminal. Their evolving conceptions of validity, spanning almost half a century and including the state-of-the-art chapters on validity in the second (Cronbach, 1971) and third (Messick, 1989a) editions of *Educational Measurement,* have provided the foundation in which other scholars have located their own work and against which alternative views of validity are compared. The other theorists whose work I cite have written repeatedly on the philosophy of validity and have proposed substantive modifications to existing conceptualizations of validity. Within the scope of this article, it is not possible to fully present the rich perspectives of these scholars or to trace their historical development. I have limited my citations to recent pieces that articulate these theorists' positions on the issues I raise. More comprehensive reviews are provided by Angoff (1988), Messick (1989a), Moss (1992), and Shepard (1993).

My intent here is to provide a primer, if you will, by highlighting major issues in validity theory and practice and by pointing readers to primary sources where these issues are addressed. I hope this will encourage debate about these issues at a crucial time when the opportunity to effect changes in practice, through the revised *Standards,* is available. As readers will note, I can't claim neutrality with respect to how the issues I raise are resolved. Although I've done my best to fairly present the range of perspectives reflected in the published literature, I've presented my own perspective as well. Moreover, I make no claim to systematically characterize assessment *practice*—the way in which the advice from validity theorists is carried out, except anecdotally or by reference to the characterization of others. In fact, one issue about what it

Pamela A. Moss is an Assistant Professor, 4220 School of Education, University of Michigan, Ann Arbor, MI 48109-1259. Her specializations are educational measurement and evaluation.

means for the *Standards* to characterize "consensus of recognized experts" is whether this refers to those who write about the philosophy of validity or to the much broader group of professionals who conduct validity research. This is an issue to which I'll return in my conclusions.

Issues

The questions I raise result from a comparative analysis of the conceptualizations of validity reflected in the 1985 *Standards* and in the recent work of validity theorists writing in the context of educational measurement. The first three questions I raise essentially address issues located within the traditional boundaries of the concept—evaluating the soundness of assessment-based interpretations—and raise issues about how to characterize the processes of conceptualizing, conducting, and reporting validity research. The second three questions focus on expanding the boundaries of validity to include consideration of the consequences of assessment use and of alternative epistemological perspectives for conducting validity research (although the answers to these questions will also influence the way in which traditional aspects of validity are carried out). For each question, I contrast the characterization of validity in the 1985 *Standards* with the varying characterizations in the work of those who write about the philosophy of validity, noting the range of alternatives proposed. Where substantial consensus exists that points to a revision, I note the features that might be included in the revised *Standards* as well as the implications for the practice of validity research; where consensus does not exist, I suggest additional issues that need to be resolved in constructing a revision.

Question 1: How Should Consensus About the Centrality of Construct Validity Be Characterized?

The authors of the 1985 *Standards,* describe validity as a unitary con-cept requiring multiple lines of evidence—including, content-, construct-, and criterion-related evidence—to support "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p. 9). There is a close to universal consensus among validity theorists—not reflected in the 1985 *Standards*—that all validity research should be guided by the principles of scientific inquiry reflected in *construct* validity. From this perspective, content- and criterion-related evidence of validity are simply two of many types of evidence that support construct validity.

When construct validity was first introduced in the 1954/1955 *Standards* (see also Cronbach & Meehl, 1955), it was viewed as an "indirect" method of validation to be used when no criterion variable or content domain could indicate the degree to which a test measures what it is intended to measure. Since then, many validity theorists have argued that no criterion variable or content domain is ever sufficient to fully justify an intended interpretation of a test score and that construct validity should form the basis of all validity research. (See Messick 1989a, 1989b, for an extended argument.)

The 1985 *Standards* moves in this direction with its unitary concept of validity requiring multiple lines of evidence but stops short of the structural reconceptualization implied in this perspective. The description of construct-related evidence, currently presented in the *Standards* as one of three types of evidence, provides a useful indicator of validity research within a construct validity framework. Essentially, it would require that validity researchers provide an *explicit conceptual or theoretical framework* to ground the intended inference and supporting evidence—not just for "psychological constructs," as indicated in the 1985 *Standards,* but for all assessment-based interpretations.

> The construct of interest for a particular test should be embedded in a conceptual framework, no matter how imperfect that framework may be. The conceptual framework specifies the meaning of the construct, distinguishes it from other constructs, and indicates how measures of the construct should relate to other variables.
>
> The process of compiling construct-related evidence for test validity starts with test development and continues until the pattern of empirical relationships between test scores and other variables clearly indicates the meaning of the test score. . . . (pp. 9–10)

Cronbach (1988, 1989), Messick (1989a, 1989b), and others who build on their work provide more elaborate descriptions of construct validity, suggesting features that might be reflected in a revision of the *Standards.* The essential purpose of construct validity is to justify a particular interpretation of a test score by *explaining* the behavior which the test score summarizes. The proposed interpretation is the construct of interest. A "strong" program of construct validation requires an explicit conceptual framework, testable hypotheses deduced from it, and multiple lines of relevant evidence to test the hypotheses. Construct validation is most efficiently guided by the testing of "plausible rival hypotheses" which suggests credible alternative explanations or meanings for the test score that are challenged and refuted by the evidence collected. "Convergent" evidence indicates that test scores are related to other measures of the same construct and to other variables that they should relate to as predicted by the conceptual framework; "discriminant" evidence indicates that test scores are not unduly related to measures of other, distinct constructs. Prominent rival hypotheses or threats to construct validity include "construct underrepresentation" and "construct-irrelevant variance" (Messick, 1989a, 1992). "Construct underrepresentation" refers to a test that is too narrow in that it fails to capture important aspects of the construct. "Construct-irrelevant variance" refers to a test that is too broad in that it requires capabilities that are irrelevant or extraneous to the proposed construct. Almost any information gathered in the process of developing and using an assessment is relevant to construct validity when it is evaluated against the theoretical rationale underlying the proposed interpretation. Thus, validation "embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated" (Messick, 1989b, p. 6). Validity conclusions are best presented in the form of an *evaluative argument* (Cronbach, 1988)

that integrates the evidence presented to justify the proposed interpretation against plausible alternative interpretations (see also Kane, 1992; Shepard, 1993).

Concerns have been raised that the "strong" program of construct validity places an excessive and impractical burden on assessment developers and users (e.g., Anastasi, 1986; Cronbach, 1989; Wiley, 1991): Cronbach notes that "the idealized strong program is most appropriate to a scientific perspective that reaches centuries into the future.... Social and behavioral scientists in particular are obligated to help their contemporaries think through their problems and evaluate proposed solutions" (p. 163). Wiley suggests distinguishing test validation from the broader concept of construct validation. Essentially, test validation examines the fit between the meaning of the test score and the measurement intent, whereas construct validation entails the evaluation of an entire theoretical framework. These concerns do not obviate the need for a program of validation research grounded in an explicit conceptual framework and articulated in an integrative argument that justifies (and refutes challenges to) the proposed meaning of the test score. The task for the authors of the revised *Standards* is to help researchers distinguish (a) what evidence is necessary to justify the use of an assessment from (b) what evidence is part of the ongoing responsibility of the measurement community at large—evidence desirable to enhance theory and practice in the long run, but beyond what can be reasonably expected of a particular developer or user.

If implemented in the revised *Standards,* this characterization of validity would not necessarily entail the collection of new or different types of evidence. It would, however, entail a more rigorous, integrative way of conceptualizing and reporting validity research. It might also entail more explicit attention to existing theory and research on similar assessments, which is now mentioned explicitly only under "validity generalization" as a fall back position when collection of local evidence is impractical. The implication for assessment developers and users is perhaps best captured by Cronbach's (1989) criticism of test manuals which "rake together miscellaneous correlations" (p. 155) rather than "report incisive checks into rival hypotheses, followed by an integrative argument" (p. 155). Within a construct validity framework, assessment developers and users would be expected to ground their research in a hypothesis generating conceptual framework, evaluating multiple lines of evidence against the conceptual framework, and presenting an evaluative argument to justify the proposed interpretation against plausible alternative interpretations.

Question 2: What Alternatives Are There to the Construct-Content-Criterion Framework for Analyzing the Concept of Validity?

The 1985 *Standards* groups the "means of accumulating validity evidence" (p. 9) into "convenient" categories of construct-, content, and criterion-related evidence of validity, which are then used to organize much of the commentary on validity. These traditional categories have been widely criticized (e.g., Anastasi, 1986, Cronbach, 1988; Messick, 1989a, 1989b; Shepard, 1993). When the concepts of content, criterion, and construct were initially used, they characterized types of validity associated with different types of inferences—from the test score to a content domain, from the test score to a criterion variable, or from the test score to a psychological construct that could not be defined by a content domain or a criterion variable. Since then, they have come to be understood as types of evidence supporting a unified notion of validity. This is reflected in the 1985 *Standards'* use of the terms content-, construct-, and criterion-*related evidence* rather than content, construct, and criterion *validity* or *validation* as they had been in previous *Standards.*

When used to categorize types of validity evidence within a unified view of validity, these traditional categories become problematic. They are neither logically distinct nor of equal importance. While content- and criterion-related evidence refer to specific types of evidence, construct-related evidence refers to every other type of relevant evidence plus content- and criterion-related evidence. Moreover, when construct validity is viewed as the basis for all validity research, it makes little sense to use the same term as one category of evidence.

While there is widespread consensus among those who write about the philosophy of validity on the inadequacy of this framework, there are no obvious alternatives here, if *obvious* is understood to mean a framework that has been widely cited and used to organize validity inquiry. (See Moss, 1992, for an overview of various category schemes that have been used for organizing presentations of validity.) Messick's (1989a, 1989b) proposed alternative to the traditional framework doesn't help in analyzing the concept of construct validity; rather, it locates construct validity in a broader notion of validity that includes explicit consideration of the value implications and consequences of assessment interpretation and use. Messick (1989a, 1989b, 1992, 1994a, 1994b) has used additional, more refined, category schemes in analyzing the concept of construct validity. One theme that reappears in many characterizations of construct validity (e.g., Anastasi, 1986; Cronbach, 1990; Messick, 1989a, 1989b), including the characterization in the 1985 *Standards,* lists illustrative types of evidence. Messick, for instance, suggests:

> We can appraise the *relevance and representativeness of the test content* in relation to the content of the domain about which inferences are to be drawn,... examine relationships among responses to the tasks, items, or parts of the test—that is, the *internal structure* of test responses, ... survey relationships of the test scores with other measures and background variables—that is, the test's *external structure,* ... directly probe the ways in which individuals cope with the items or tasks, in an effort to illuminate the *processes underlying item response and task performance,* ... investigate *uniformities and differences in these test processes and structures over time or across groups and settings*—that is, the generalizability of test interpretation and use, ... see if the test scores display appropriate *variations as a function of instructional*

> *and other interventions,* . . . [and, in the broader view of validity] appraise the *value implications and social consequences* of interpreting and using the test scores. (Italics added; 1989b, p. 6)

More recently, Messick (1994a, 1994b) has highlighted a different category scheme from his chapter on validity (1989a). Messick's "aspects of construct validity" (1994b, p. 8) can be mapped onto his categories of evidence described above; however, he characterizes them as "general validity criteria or standards" (1994b, p. 8).

> The *content* aspect of construct validity includes evidence of content relevance, representativeness, and technical quality. . . .
>
> The *substantive* aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance. . . , along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessments tasks.
>
> The *structural* aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue. . . .
>
> The *generalizability* aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks . . . , including validity generalization of test-criterion relationships. . . .
>
> The *external* aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons . . . , as well as evidence of criterion relevance and applied utility. . . .
>
> The *consequential* aspect [included in a broader view of validity discussed below] appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice. (Italics added; Messick, 1994b, p. 9)

There are a number of closely related issues that need to be addressed in analyzing the concept of validity into categories to guide validity research, including the issue of whether general categories should be used at all. First, at what level of generality should categories be presented—(a) as relevant for all validity research, as the traditional categories now are; (b) for different purposes and types of assessment; or (c) only for specific cases of assessment use in specific contexts? Some have argued that *specific cases* or *exemplars* of validity research would better illustrate the concept (e.g., Mishler, 1990), because abstract categories can be insensitive to context specific variations. At the level of purpose, Shepard (cited in Kane, 1992) suggests that validity research proceed by articulating the assumptions necessary to justify the use of a test for a particular purpose and by collecting evidence to support or challenge those assumptions. She then illustrates this with specific cases of validity research. These alternatives are not necessarily mutually exclusive—for instance, illustrative categories of evidence could be used at the general level, followed by more specific categories at the level of purpose, illustrated, as Shepard has, with cases of validity research.

Second, if categories are used, should they be perceived as examples of types of evidence, or should they be perceived as "mutually exclusive, exhaustive of the possible lines of evidence . . . , and mandatory" (Loevinger, 1957, pp. 653–654)? Mandatory categories may not be responsive to particular situations and may constrain practice (see Question 6 below); illustrative categories rely on professional judgment and risk encouraging an anything-goes mentality. The 1985 *Standards* adopts an intermediate position with respect to this issue, asserting that an ideal validation spans all three traditional categories but that a single line of solid evidence is preferable to numerous lines of questionable quality. On the other hand, consideration of each of Messick's aspects of validity appears more mandatory, because he advises that "evidence pertinent to all of these aspects needs to be integrated into an overall validity judgment . . . , or else . . . compelling reasons [provided] why not" (1994b, p. 15). Third, should categories describe types of evidence, as the 1985 *Standards* do, or other aspects of validity inquiry? For instance, others have proposed schemes that categorize (a) specific inferences or assumptions necessary to justify a proposed use (e.g., Kane, 1992; Shepard, 1993); (b) the activities undertaken at different stages of test development and use, beginning with a definition of the construct (e.g., Cole & Moss, 1989; Haertel, 1985); or (c) the criteria by which evidence should be evaluated, such as cognitive complexity or content quality (e.g., Linn, Baker, & Dunbar, 1991). (These three issues will also be relevant in considering how to analyze a broader conception of validity.)

The concern might be raised that to abandon such venerable and widely used concepts as the construct-content-criterion categories risks confusion. This is a concern that cannot be taken lightly. Shepard (1993) counters that the practice of retaining the traditional names while changing their meaning (from types of validity to types of evidence) has not highlighted the significance of the change in meaning: "Because existing terminology has been imbued with new meanings (rather than inventing new terms to signify changed understandings), it is possible for students of measurement to persist in the old forms" (p. 407). Certainly the transition could be eased with a careful mapping or translation of any new category labels in terms of the old. The goal here, as for the 1985 *Standards,* is to clearly communicate the features of sound validity research as well as to "assure that relevant issues are addressed" (p. 2).

Question 3: How Should the Relationship Between Validity and Other Concepts Associated With Test Evaluation Be Articulated?

If almost any evidence gathered during the development and evaluation of an assessment is relevant to construct validity, then it becomes important to articulate the relationship between validity and these various types of evidence. Here, I focus on the concepts of reliability, bias, and fairness, although the issue is relevant to other concepts addressed in the *Standards.*

Reliability is presented in the 1985 *Standards* as a concept distinct from

validity. Moreover, the relationship between reliability and validity is not explicitly articulated. *Reliability* is defined as freedom from errors of measurement and as consistency among measures intended as interchangeable. With more complex assessments, reflecting integration of multiple skills and knowledge that may vary from task to task, distinctions between reliability and construct validity blur. This is because with complex assessments it becomes harder to distinguish "interchangeable" measures from different measures of the same construct (Moss, 1994; Wiley & Haertel, in press). Whether two measures should be treated as interchangeable or not is a matter of theoretical choice that requires logical and empirical justification—a choice which should be integrated into the construct validity argument. Some theorists have incorporated reliability/generalizability into the concept of validity, including it as one of their analytic categories (e.g., Haertel, 1985; Linn, Baker, & Dunbar, 1991; Messick, 1994a, 1994b). Moreover, given the often noted tension between reliability and complexity or authenticity with performance assessment, some have noted the importance of balancing these concerns in reaching and justifying an overall validity conclusion (Linn, Baker, & Dunbar, 1991; Messick, 1994a). Again, the rationale needs to be articulated in the overall validity argument.

Similar concerns arise about the relationship between validity, bias, and fairness. The only mention of bias in the section on validity is in the context of differential prediction and criterion-related evidence. Cole and Moss (1989) define *bias* as "differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (p. 205), thus making explicit the relevance of bias concerns to any kind of evidence impinging on the interpretation of test scores. While I doubt that the authors of the 1985 *Standards* would disagree with this definition, their reference to bias in the chapter on validity is narrow and ad hoc. With respect to fairness, the 1985 *Standards* carefully delimits concerns about validity and bias from concerns about the broader concept of fairness, which falls outside the guidance they provide: "fairness is not a technical psychometric term; it is subject to different definitions in different social and political circumstances" (p. 13). How the revised *Standards* might address the relationship between validity, evaluation of consequences, and fairness (which encompasses concerns about differential interpretive validity as well as differential impact) is a controversial set of issues to which I turn next.

Question 4: To What Extent Should the Standards *Foreground Use Over Interpretation in Defining (Construct) Validity?*

With this question, I move beyond the traditional validity focus on interpretation to consider issues associated with the consequences of assessment. With Questions 4 and 5, I distinguish two levels at which consequences might be considered. Question 4 addresses expanding the concept of validity to include consideration of the consequences associated with the immediate and expressed purposes of assessment (e.g., placement, selection, program evaluation, etc.). Question 5 focuses on intended and unintended consequences beyond those expressed in the immediate purposes of testing (such as, concerns about differential impact associated with fairness, concerns about enhancing or narrowing instruction, etc.).

The 1985 *Standards* gives primacy to inferences or interpretations over uses in the definition of validity: "Validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p. 9). The language of many of the associated standards reinforces this emphasis. Messick (1989a, 1989b) highlights the importance of investigating the validity of a proposed use by distinguishing use from interpretation in his analysis of validity. Construct validity, which encompasses the "evidential basis of test interpretation," addresses the soundness of an inference, without reference to a particular use, and "construct validity plus relevance/utility," which encompasses the "evidential basis of test use," refers to the appropriateness of an interpretation for a given context and use. He then distinguishes the evidential basis of test interpretation and use which, like the 1985 *Standards*, focuses on inferences from the "consequential basis of test interpretation and use" which focuses on value implications and on actual and potential outcomes. Shepard (1993) criticizes these distinctions between interpretation and use and between evidence and consequence, although not the importance of the issues they highlight. As she notes, the distinctions imply that construct validity would initially proceed in the same way, regardless of how a measure is used, merely testing additional hypotheses for the applied purpose. This approach "would be acceptable if researchers had infinite resources to test . . . all possible theoretical and practical relationships" (p. 429), but it does not help applied researchers prioritize validity questions. Moreover, it appears to perpetuate a distinction between facts and values that Messick does not intend.

In contrast, Shepard (1993), citing Kane (1992), argues that construct validity should be guided by questions of use: "What does the test claim to do?" (Shepard, 1993, p. 429). At the level of language, this point may seem subtle; however, at the level of practice, it may have substantial implications for kinds of evidence required to justify the continued use of a test. Shepard (1993) argues, for instance, that when a test is used for placement evidence is needed about whether students actually benefit from the differential instruction; mere predictive correlations would not be sufficient. She notes that this focus on particular uses in guiding validity research serves a pragmatic purpose as well—by helping researchers set priorities in addressing validity questions most relevant to the context of use. When we evaluate the use of a test for a given purpose, we expand the necessary kinds of evidence about how the assessment works within the system in which it is used (Cronbach, 1980). This leads into the broader question of the role of consequences in validity research.

Question 5: To What Extent Should the Standards *Incorporate the Investigation of Consequences Into Its Conceptualization of Validity?*

Many measurement theorists have argued for the importance of expanding the concept of validity to include explicit consideration of intended and unintended consequences of assessment use (e.g., Cronbach, 1988; Haertel, 1992; Linn, 1993; Messick, 1989a; Moss, 1992; Shepard, 1993). The question about refocusing validity to foreground use is a piece of this issue. Here the issue is how far to expand the boundaries of validity and/or the associated responsibilities for assessment developers and users. While there is little dispute about the significance of consequences, there are at least three distinguishable issues to face in revising the *Standards:* (a) whether or not the *Standards* should encourage or require assessment developers and users to consider evidence about consequences; (b) to what extent the consideration should address *actual* consequences, thus requiring evidence about the outcomes of assessment use, or *potential* consequences, thus requiring careful hypothesizing and use of existing research; and (c) whether that consideration of consequences should be viewed as an aspect of validity or as a distinct concept.

In the validity chapter, the 1985 *Standards* is, for the most part, silent on investigating consequences of test use, except in certain specific and limited circumstances. While one might interpret the word *appropriateness* in the definition of validity to cover consequences ("Validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores," p. 9), there is little in the specific validity standards to support such an interpretation. As I noted above, technical issues of bias are delimited from the social and political concerns of fairness, which fall outside the guidance provided. There is one standard in this chapter that refers to evidence about outcomes for placement purposes—"evidence of a test's differential prediction for . . . [classification into alternative treatment groups] should be provided" (p. 18)—but it is identified as "secondary" versus "primary" (p. 18) implying that it is desirable, but not required. References to consequences do appear occasionally outside the validity chapter in the chapters on test use. For instance, a primary standard for general use states, "Test users should be alert to probable unintended consequences of test use and should attempt to avoid actions that have unintended negative consequences" (p. 42).

Cronbach (1988), Messick (1989a), and others (e.g., Haertel, 1992; Linn, 1993; Shepard, 1993) present a much more comprehensive view of the meaning of validity and the responsibilities of validity researchers. Messick defines *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" (p. 13), thus building consideration of the value implications and social consequences of test interpretation and use into his conception of validity. Cronbach (1988) argues that validity argument "must link concepts, evidence, social and personal consequences, and values" (p. 4). Haertel (1992), Linn (1993), and Shepard (1993) draw on this work in characterizing the concept of validity.

A few writers have raised concerns about overburdening the concept of validity to the point where it ceases to provide useful guidance. Wiley, for instance, argued:

> A valid set of measurements—defined in terms of realized intent—may be badly used. . . .The understanding of these *use errors* is conceptually and socially important, but involves social and moral analyses beyond the scope of test validation as defined here and would needlessly complicate the conception and definition of test validity. (1991, p. 88)

This suggests treating the evaluation of consequences as a concept distinct from validity.

Others have argued, in different but complementary ways, that the concepts are not distinct. Messick (1989a, 1989b), for instance, argues that we should expand the meaning of validity because the consequences of test interpretation and use are "signs of validity or invalidity" (1989b, p.11) and thus integrally related to score meaning: "A social consequence of testing either stems from a source of test *in*validity or else reflects a valid property of the construct assessed, or both" (p. 10). Extending this approach, Shepard (1993) argues that unintended consequences are simply rival hypotheses to the expressed purpose of testing and so should be considered part of the validity argument: "Pursuing unintended effects is a logical extension of [the] inclusion of rival hypotheses when framing validity evaluations" (p. 426). Cronbach (1988) argues that expanding the concept of validity appropriately acknowledges the crucial importance of guarding against adverse social consequences: "The bottom line is that validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences. . . . You . . . may prefer to exclude reflection on consequences from the meanings of the word validation, but you cannot deny the obligation" (p. 6). Sometimes the expressed intent of assessment is to reform the system in which an assessment is used (Linn, 1993), and so measurement intents and anticipated consequences merge.

The authors of the 1985 *Standards* raised the concern that the document not become "a social action prescription" (p. v). This concern could be addressed by careful attention to another expressed goal: "The necessary technical information [should] be made available so that those involved in policy debate may be fully informed" (p. 1). To expect that assessment developers and/or users will make available evidence about the potential or actual consequences in order to inform policy debate is not the same as prescribing social policy—indeed, it is little different from requiring that they make available other types of evidence to inform judgment about the soundness of the interpretation. Moreover, no aspect of validity research is value free: as Cronbach (1980) notes, "the statement that a test producer ought to examine certain factual questions is value-laden, and failure to call for investigation of certain others is in itself taking sides on political matters" (p. 105).

There is a real and legitimate concern about the cost of requiring assessment developers and/or users to

provide additional evidence and of possibly delaying or precluding the operational use of a potentially beneficial assessment. However, not to consider such evidence risks either unrealized intentions or unintended negative consequences (Haertel, 1992; Linn, 1993). Haertel, for instance, advises:

> The consequences of performance measurement . . . merit continual attention. The expected benefits . . . may or may not materialize, and negative side effects may or may not occur. For instance, student performance testing may encourage better instructional approaches . . . or may lead to an undesirable narrowing of the curriculum. . . . New testing formats may bring diminution in racial, ethnic, or gender group differences, or may exaggerate those differences. (p. 988)

Again, this issue might be resolved by providing sound and practical advice that helps developers and users distinguish (a) what evidence is necessary before the first operational use of an assessment from (b) what evidence is necessary to justify its continued use and from (c) what evidence is the ongoing responsibility of the measurement community at large.

Here, as with categories of evidence for construct validity, it might be appropriate to distinguish between advice provided at the general level—for all purposes of assessment—and advice provided at the level of particular purposes. Messick (1989a), Cronbach (1988, 1989), and Shepard (1993) offer general advice for evaluating the intended and unintended consequences of a test interpretation and use. To evaluate the potential consequences of a test use, Messick suggests pitting the proposed use against alternative assessment techniques and alternative means of serving the same purpose, including the generalized alternative of not assessing at all. Cronbach (1988, 1989) articulates distinctions among functional, political, and economic consequences of assessment. He suggests using stakeholders' interests as well as evaluators' concerns to generate a list of potential questions and then prioritizing the questions based on (a) prior uncertainty about the issue, (b) information to be yielded by a feasible study compared to how much uncertainty will remain, (c) cost of the investigation in terms of time and dollars, and (d) leverage for achieving consensus about the use of the test in the relevant audience. Shepard (1993) suggests using the intended purposes of assessment to generate rival hypotheses about unintended consequences and then prioritizing questions in light of the seriousness of consequences for individuals and programs.

For many purposes of assessment, there is a history of accumulated research into consequences of particular concern (e.g., about the impact of high-stakes assessment on teaching or learning [Linn, 1993; Linn, Baker, & Dunbar, 1991], about the impact of different selection models on various groups of concern [e.g., Cole & Moss, 1989], etc.). The authors of the *Standards* might draw on the accumulated experience of research tradition with various purposes to suggest important categories of consequential evidence to consider with respect to those different purposes. At the very least, assessment developers might be expected to address *potential* consequences by summarizing the existing evidence of using the assessments like the one in question for the proposed purpose and context. And, when an anticipated consequence of assessment is explicit in the purpose of assessment (e.g., to raise educational standards) or when the anticipated consequences "impinge on the rights and life chances of individuals" (Cronbach, 1988, p. 6) (e.g., to certify for high school graduation), the investigation of consequences becomes particularly salient.

Question 6: Should the Standards *Incorporate Principles From Research Traditions Other Than Psychometrics in Characterizing Validity to Support the Evaluation of Less Standardized Assessments?*

Given the goal of the *Standards* to characterize existing consensus among measurement experts, the answer to this question is "no:" Few measurement theorists have addressed the issue explicitly. However, the issue cannot be ignored. There are already assessment practices in use for high-stakes purposes that do not (perhaps cannot) yield the type of evidence expected in mainstream approaches to validity research. Consider, for instance, the practice of some schools where certification for graduation is like a dissertation exam. Students, in negotiation with teachers, prepare one or more exhibits of their work, and a committee meets to debate and evaluate the merits of that work (e.g., Darling-Hammond & Snyder, 1992). This purpose of certification for graduation falls within the purview of the *Standards,* but this format does not. Evidence of reliability/generalizability across readers and tasks—at least in the way these concepts are typically operationalized—is not obtainable because the tasks are not evaluated independently. In fact, to require such evidence would substantially alter the nature of the assessment and, some would argue, decrease its validity. Moss (1994) and others (e.g., Johnston, 1989) have suggested using validity principles from interpretive research traditions to assist in evaluating these and other less standardized assessment practices, thereby exploring alternative means for serving the important epistemological and ethical purposes that underlie traditional validity practices.

Among members of the psychometric community, Messick opened the door here with his (1989a) discussion of a "Singerian" mode of inquiry: He suggests observing or evaluating one inquiring system in terms of another to probe the methodological and value assumptions underlying each system. Drawing on this advice, Moss (1994) suggests contrasting psychometric and interpretive (especially hermeneutic) approaches for drawing and evaluating assessment-based interpretations to highlight assumptions and consequences to teachers and students of using more standardized forms of assessment.

This is an issue that will require careful consideration. As the above example illustrates, current validity theory discourages such practices by requiring evidence that may not be possible to provide. Should the *Standards* continue to require evidence that discourages these developments? Should the *Standards* venture into relatively uncharted

territory and attempt to provide guidance for investigating the validity of such nonstandardized assessments (guidance which may fall well outside the accepted principles of psychometrics)? Should the *Standards* delimit its applicability to acknowledge that certain purposes for assessment may be served by assessment formats that are not themselves covered by the *Standards*? If this strategy is adopted, does it open the door for assessment developers and users to argue that their assessments, too, fall outside the formats covered, thus weakening the ability of the *Standards* to encourage sound professional practice? There are no easy answers.

Concluding Comments

The disjunctions in validity theory that I've summarized, between the 1985 *Standards* and the work of many measurement scholars, already existed when the 1985 *Standards* was published. In fact, much the same article could have been written over a decade ago, with minor changes in dates and names (and without the question on alternative epistemological perspectives, which was not prominent in the discourse among members of the measurement community). By 1980, there was already an emerging consensus among validity theorists about the inadequacy of the construct-content-criterion framework for guiding validity research, about the centrality of construct validity to the evaluation of any assessment-based interpretation, and about the importance of expanding the concept of validity to include explicit consideration of the consequences of assessment use. (For reviews, see Angoff, 1988; Messick, 1989a; Moss, 1992; Shepard, 1993.)

In characterizing expert consensus, the choice to focus on validity theorists rather than on the larger group of professionals who develop, use, and evaluate assessments is a controversial one. Cronbach (1988), Messick (1989a), and Shepard (1993) all comment on the gap between current validity theory and the practice of much validity research. If the consensus reflected in the revised *Standards* is faithful to the practice of validity research, then revisions are likely to be far less extensive. This would be, I believe, a mistake. The fact is that the *Standards* not only reflect consensus, they shape it—both directly, in the practices they promote (and discourage), and, indirectly, through textbooks used in educating assessment professionals. Although, as Messick (1989a) notes, it is not an inappropriate role for professional standards to codify sound practice, "the price paid in the politics of compromise is that the standards downplay principles that would lead the measurement field forward—that is, the formulation of desirable, though challenging, next steps toward improved testing practice" (p. 92). If the revised *Standards* reflects largely the practices that the existing *Standards* shapes, then it is hard to imagine how the inertia of tradition will be overcome.

Toulmin (1972), a historian and philosopher of science, comments on possibilities for engaging in rational conceptual change within a discipline. He suggests that the task of choosing among alternative conceptual frameworks amounts to making a prediction, a "rational bet" (p. 487), about which set of concepts will better fulfill the long-term ambitions of the discipline involved. The rationality of that prediction is based on an appraisal of past experience with conceptual changes in terms of their success in furthering the goals of the discipline and their substantive relevance to the current situation.

This is useful advice. There is much we can learn from our past experience with various assessment practices and the principles by which they have been evaluated. Consider, for instance, what we have learned about the role that multiple-choice tests can play when used for high-stakes purposes, in narrowing the curriculum, and possibly resulting in differential access to opportunities to learn for students in low scoring schools (Linn, 1993). Such tests were in operational use, with associated rewards and sanctions, long before systematic evidence was available about their consequences for teaching and learning. What might we have done differently if providing evidence about consequences had been considered a necessary part of sound professional practice? Linn raises the concern that the nation is about to embark on another test-based reform, now with high optimism about the beneficial effects of performance assessment, but again without adequate quality control to evaluate the soundness of those assumptions. What can we learn from our past experience to better inform these crucial policy decisions?

I hope members of NCME will participate actively in the dialogue surrounding the revision of the *Standards*. Comments for the joint committee revising the *Standards* can be sent to Dan Eignor, NCME's liaison to the committee, at Educational Testing Service, Princeton, NJ, 08541. The revision of the *Standards* provides a rare—once per decade—opportunity for us to reconsider the guiding principles of our profession at a time when we can make changes that have a substantial, positive impact on the practice of validity research and on the community of stakeholders we serve. Let's take advantage of it.

Note

I am grateful to Lorrie Shepard for her detailed and thoughtful comments on an earlier draft of this article, to members of the Joint Committee on the *Standards for Educational and Psychological Testing* who continually inform and challenge my perspective on these issues, and to the National Academy of Education/Spencer Post-Doctoral Fellowship Program for providing me with the time to complete this work. The views expressed in this article are my own; they are not intended to represent the views of the joint committee or any of its members.

References

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

American Educational Research Association, & National Council on Measurements Used in Education. (1955). *Technical recommendations for achievement tests*. Washington, DC: National Education Association.

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

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. 19–32). Hillsdale, NJ: Erlbaum.

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

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. (1980). Validity on parole: How can we go straight? *New directions for testing and measurement: Measuring achievement, progress over a decade* (No. 5, pp. 99–108). San Francisco: Jossey-Bass.

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

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

Cronbach, L. J. (1990). *Essentials of psychological testing.* New York: Harper & Row.

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

Darling-Hammond, L., & Snyder, J. (1992). Reframing accountability: Creating learner-centered schools. In A. Lieberman (Ed.), *The changing contexts of teaching* (Ninety-First Yearbook of the National Society for the Study of Education (pp. 11–36). Chicago: University of Chicago Press.

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research, 55,* 23–46.

Haertel, E. (1991). New forms of teacher assessment. *Review of Research in Education, 17,* 3–29.

Haertel, E. (1992). Performance measurement. In *Encyclopedia of educational research* (6th ed., pp. 984–989). New York: Macmillan.

Johnston, P. (1989). Constructive evaluation and the improvement of teaching and learning. *Teachers College Record, 90* (4), 509–528.

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

Landy, F. J. (1986). Stamp collecting versus science. *American Psychologist, 41,* 1183–1192.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15,* 1–16.

Linn, R. L. Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), 5–21.

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

Messick, S. (1989a). Validity. In. R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education & National Council on Measurement in Education.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18* (2), 5–11.

Messick, S. (1992). Validity of test interpretation and use. In *Encyclopedia of educational research* (6th ed., pp. 1487–1495). New York: Macmillan.

Messick, S. (1994a). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23* (2), 13–23.

Messick, S. (1994b, June). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.* Keynote address presented at the Conference on Contemporary Psychological Assessment, Stockholm, Sweden.

Mishler, E. G. (1990). Validation in inquiry-guided research. *Harvard Educational Review, 60,* 415–442.

Moss, P. A. (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* (2), 5–12.

Shepard, L. A. (1993). Evaluating test validity. In *Review of Research in Education, 19,* 405–450.

Toulmin, S. (1972). Human understanding. Princeton, NJ: Princeton University Press.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach* (pp. 75–107). Hillsdale, NJ: Erlbaum.

Wiley, D. E., & Haertel, E. H. (in press). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In R. Mitchell & M. Kane (Eds.), *Implementing performance assessment: Promises, problems, and challenges.* Washington, DC: Pelavin.