Psychological Methods
2012, Vol. 17, No. 1, 31–43

1082-989X/12/$12.00 DOI: 10.1037/a0026975

Defining and Distinguishing Validity: Interpretations of Score Meaning and Justifications of Test Use

Gregory J. Cizek
University of North Carolina at Chapel Hill

The concept of *validity* has suffered because the term has been used to refer to 2 incompatible concerns: the degree of support for specified interpretations of test scores (i.e., intended score meaning) and the degree of support for specified applications (i.e., intended test uses). This article has 3 purposes: (a) to provide a brief summary of current validity theory, (b) to illustrate the incompatibility of incorporating score meaning and score use into a single concept, and (c) to propose and describe a framework that both accommodates and differentiates validation of test score inferences and justification of test use.

Keywords: validity, validation

Measurement instruments are integral to training, practice, and research in the social sciences. The quality of the data yielded by those instruments is an essential concern of measurement specialists, test users, and consumers of test information. That concern is heightened whenever test results inform important decisions including, for example, judging the effectiveness of interventions; awarding credentials, licenses, or diplomas; making personnel decisions; and countless other situations in which the information yielded by a test has meaningful consequences.

Professional standards for these instruments have existed for over 50 years, beginning with the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (American Psychological Association, 1954) and spanning five editions to the current *Standards for Educational and Psychological Testing* (hereafter, *Standards*; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999). Despite changes in each revision of the *Standards*, common themes are evident, particularly regarding the dependability (i.e., reliability) and meaning (i.e., validity) of test-generated data.

One topic—validity—has evolved appreciably (Geisinger, 1992), but the primacy of that characteristic has been consistently affirmed. In 1961, Ebel referred to validity as "one of the major deities in the pantheon of the psychometrician" (p. 640). Nearly 40 years later, the current *Standards* describe validity to be "the most fundamental consideration in developing and evaluating tests" (AERA, APA, & NCME, 1999, p. 9). In his 2005 Presidential Address to the National Council on Measurement in Education, Frisbie asserted that validity is "a concept that is the foundation for virtually all of our measurement work" (p. 21).

Despite its honored status, all is not well with validity. This article is organized into four sections that (a) provide a brief introduction to contemporary validity theory; (b) describe a lingering concern regarding the concept of validity; (c) propose a comprehensive remedy to address the concern; and (d) suggest implications for further theoretical development of validity theory, professional standards, and validation practice.

Background on Contemporary Validity Theory

There would seem to be broad consensus about several aspects of contemporary validity theory. A brief summary of these points of agreement is presented in this section.

A first tenet of contemporary validity theory is that validity pertains to the intended inferences or interpretations made from test scores (Kane, 2006; Messick, 1989). Relatedly, it is the intended score inferences that are validated, not the test itself. The current *Standards* indicate that "it is the interpretations of test scores that are evaluated, not the test itself" (APA, AERA, & NCME, 1999, p. 9; see also Cronbach, 1971). The concept of score interpretation or inference is central to current thinking about validity. Because latent characteristics, traits, abilities, and so on cannot be directly observed, they must be studied indirectly via the instruments developed to measure them. Inference is required whenever it is desired to use the observed measurements as an indication of standing on the unobservable characteristic. Because validity applies to the inferences made from test scores, it follows that a clear statement of the intended inferences is necessary to design and conduct validation efforts.

A second point of consensus is that the notion of discrete kinds of validity (i.e., content, criterion, construct) referred to as the "trinitarian" view (Guion, 1980, p. 385) has been supplanted by what has been referred to as the unified view of validity. The unified view posits that all evidence that might be brought to bear in support of an intended inference is evidence bearing on the

This article was published Online First January 23, 2012.

Correspondence concerning this article should be addressed to Gregory J. Cizek, School of Education, Program in Educational Psychology, Measurement, and Evaluation, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3500. E-mail: cizek@unc.edu

construct the instrument purports to measure.[1] The unitary perspective was foreshadowed over 40 years ago by Loevinger, who observed 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" (1957, p. 636); the unitary perspective was subsequently incorporated into the *Standards* (see APA, AERA, & NCME, 1974, p. 26).

A third point of consensus is that validity is not expressed as the presence or absence of that characteristic. According to Zumbo, "validity statements are not dichotomous (valid/invalid) but rather are described on a continuum" (2007, p. 50). That continuum is one along which those who evaluate validity evidence arrive at judgments about the extent to which the corpus of accumulated evidence supports the intended test score inference. Those judgments are necessarily integrative and evaluative for at least three reasons. For one, synthesis must occur if more than only one source of evidence is mined; searching validation efforts typically tap multiple evidentiary sources. For another, these collections of evidence are typically mixed in terms of how directly each piece of evidence bears on the intended inference, the weight given to the various sources of evidence will vary, and the degree of support for the intended inference contributed by the various sources of evidence will differ. Finally, validation necessarily involves the application of values (e.g., regarding which sources of evidence should be admitted, the relevance of those sources, how they should be weighted, the favorability/unfavorability of the evidence). Even the most thorough validation efforts can yield equivocal evidence that can lead equally qualified evaluators to different conclusions depending on the tacit or explicit background beliefs and assumptions that affect their perceptions of the evidential value of the information to be synthesized (see Longino, 1990).

A final point of agreement is that validation is an ongoing endeavor. For example, summarizing "every . . . treatise on the topic," Shepard has observed that "construct validation is a never-ending process" (1993, p. 407). Many factors necessitate a continuing review of the information that undergirds an intended test score inference. Such factors include accumulated experience with administration and scoring of the instrument, the availability of previously unknown or unavailable sources of evidence, replications of the initial validation effort, information from administration of the instrument in new contexts or populations, and theoretical evolution of the construct itself—any one of which could alter the original judgments about the strength of the validity case.

A Central Problem in Contemporary Validity Theory

Despite broad endorsement of the importance of validity and agreement on many tenets of modern validity theory, there are also areas of disagreement and areas of concern including, for example, the philosophical foundations of the concept (see Hood, 2009) and the boundaries of the concept (see Borsboom, Mellenbergh, & van Heerden, 2004). A particularly problematic concern is the very definition of *validity*.

It would seem desirable to begin consideration of validity by providing a clear, broadly accepted definition of the term. Unfortunately, a straightforward definition of validity is not found in the contemporary psychometric literature. For example, none of the 14 chapters in the edited book *Test Validity* (Wainer & Braun, 1988) provides a definition of *validity*, nor does the chapter by Kane (2006) in the current edition of *Educational Measurement* (although a definition of *validation* is provided). Perhaps the most well known description of validity is found in Messick's (1989) chapter in the third edition of *Educational Measurement*, where the concept is defined 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). However, as others have also observed, that definition conflates validity (a property) with validation (a process; see, e.g., Borsboom et al., 2004); Hood (2009) has also critiqued Messick's explication of the concept as containing "three mutually inconsistent accounts of what 'validity' means" (p. 458).

As will be demonstrated, a clear definition of *validity* is possible, and definitions of both *validity* and *validation* are provided subsequently. To begin a consideration of reconceptualizing validity, however, further examination of the current state of thinking about the term *validity* is necessary. That a clear definition of validity does not exist is a deficiency but also a symptom of a central problem in current validity theory. Namely, the very concept of validity has been conceptualized in a way that actually prevents clarity in definition—with the ultimate and unfortunate result of deterring alacrity in the practice of validation.

Roots of the Problem

Although the roots of the current concern can be traced to, for example, notions introduced by Cronbach (1971) and others, the concern is perhaps most visible in the familiar conceptualization of validity put forth by Samuel Messick, who is perhaps most widely recognized as a leading theorist and explicator of modern validity theory. In a 1980 article (and subsequently in a chapter in the third edition of *Educational Measurement*) Messick introduced a matrix that formalized validity as comprising a constellation of what he referred to as four "facets of validity" (1989, p. 20).

The four facets of the matrix represent the intersections that result from cutting validity in two ways. One half of the matrix can be seen in its rows, where validity is separated into an evidential basis (the top row) and a consequential basis (the bottom row). The other half of the matrix separates validity concerns into two columns that address test score interpretation (on the left) and test score use (on the right). To be sure, Messick's work in general and the matrix in particular have stimulated much thinking, debate, and attention to the concept of validity. Indeed, somewhat of a cottage industry has been engaged in interpreting, deconstructing, reconstructing, clarifying, and elaborating on the seminal ideas Messick introduced.

Although apparently providing a crisp heuristic for considering validity, the relatively simple 2×2 matrix has actually engendered some enduring controversies and theoretical difficulties. For example, one of the intersections of the matrix (what has been called the consequential basis of test use, or *consequential validity* as shorthand) has been a point of controversy since at least 1966, when Tenopyr commented that "to speak of 'consequential valid-

[1] Although the unitary view of validity is currently the most widely accepted, it is not universally endorsed; see, for example, Borsboom (2005, Chapter 6).

ity' is a perversion of the scientific underpinnings of measurement" (p. 14). Since that time, the problem of incorporating social consequences of test use as a source of validity evidence has been a lingering source of controversy and spawned considerable debate (see, e.g., Cizek, 2011; Hubley & Zumbo, 2011; Mehrens, 1997; Moss, 1998; Popham, 1997). Prototypical examples attempting to (re)interpret or deconstruct Messick's work are found, for example, in Hubley and Zumbo, who argue that "most [measurement specialists] do not know about, do not understand, or have misinterpreted Messick's characterization of consequences" (2011, p. 220), and in Markus (1998), who observed that "questioning Messick's theory of validity is akin to carving a Thanksgiving armadillo" (p. 7).

The difficulties related to interpreting Messick's work and the dust-ups related to consequences are secondary, however, to a more fundamental concern for the concept of validity that is inherent in the theoretical framework captured, at least in part, by the matrix. Namely, the matrix and accompanying theoretical explication laden the term *validity* with more than a single term can bear. Despite broad agreement about the importance of validity and the major tenets of modern validity theory described previously, disagreement exists regarding the very definition and boundaries of the concept.

Validity's troubles are both philosophical and practical. For example, Hood (2009) has explored the philosophical foundations of the competing validity perspectives of Messick (1989) and Borsboom et al. (2004). Hood suggested that "these two approaches are potentially complementary" (p. 453) but concluded that the competing and unresolved philosophical tensions position validity as "psychology's measurement problem" (p. 451). The more practical problem of defining *validity* is the primary concern addressed in this article although, as will be seen, addressing the practical problem necessarily suggests theoretical refinements.

Validity: Unified or Fragmented?

In contrast to the broadly accepted notion that validity is a unitary concept, the four facets of the matrix actually portray four highly distinct and important testing concerns. Although each cell in the matrix is a valuable aspect of sound testing practice, they cannot be subsumed under the single concept of *validity*.

For ease of reference, consider the four cells of the matrix following traditional cell nomenclature. Accordingly, the four cells of the matrix can be identified as

Cell A: Evidential basis of test interpretation,

Cell B: Evidential basis of test use,

Cell C: Consequential basis of test interpretation, and

Cell D: Consequential basis of test use.

Considering the four cells in reverse order, the row containing Cells C and D is identified as the consequential basis. As it turns out, this row is perhaps the least controversial—although not without conceptual and practical difficulties. Cell D highlights a concern about the social consequences of test use. Messick defined this facet as addressing "the functional worth of scores in terms of social consequences and their use" (1989, p. 84). According to Messick, this facet of validity evidence requires "an appraisal of the potential social consequences of the proposed use and of the actual consequences when used" (1980, p. 1023), a view that has come to be referred to in abbreviated fashion as *consequential validity*.[2] On the one hand, it is the incorporation of this component *as a facet of validity* and the inclusion of evidence based on consequences of testing *as a source of validity evidence* that has sparked continuing debate about so-called consequential validity. On the other hand, there would seem to be little disagreement that the social consequences of test use are a central concern of, at minimum, test users and arguably a concern of all persons involved in or affected by testing and that the social consequences of testing should be accounted for in a comprehensive approach to professional testing practice.

Cell C, at the intersection of consequential basis and test interpretation, highlights the proposition that value considerations underlie score meaning and are inherent at every stage of test development, administration, and reporting, beginning with the very conceptualization of a test. According to Messick (1989, 1998), the consequential basis of test interpretation requires explicit consideration of the social values that influence theoretical interpretations of test scores, the meanings ascribed to test score interpretations, the questions that tests are constructed to answer, and how they go about doing so. His assertion that such issues are not typically vigorously addressed may be equally true today. Indeed, as Messick has stated: "Value considerations impinge upon measurement in a variety of ways, but especially when a choice is made in a particular instance to measure some things and not others" (1975, p. 962). However, as will be argued shortly, the difficulty here is that the social consequences of test use do not bear on the validity of score interpretations. This fact is perhaps primarily responsible for the impossibility of producing a coherent, stand-alone definition of *validity*.

Cell B indicates that attention should be paid to the evidential basis for test use. Messick indicated that the evidentiary basis includes "construct validity + relevance/utility" (1989, p. 20). Here also some relevant concerns are explicit, though entangled. Construct validity—identified as the single entry in Cell A—is a necessary, though not sufficient, requirement for the defensible development and use of a measure. However, relevance and utility of test scores is a distinguishable concern—indeed, a concern that not only can but must be distinguished from score interpretation in a comprehensive theoretical framework to guide test development and use.

Finally, Cell A focuses on the evidential basis for test (score) interpretation. As just mentioned, the label in this cell of the matrix is simply *construct validity*. The positioning of construct validity here reflects the well-accepted unitary view in which, in shorthand expression, all validity is construct validity. To the extent that this view is reasonable, Cell A captures what is nearly the whole of validity. That is, if the concept of validity is limited to the degree of support for test score meaning with respect to a test-taker's standing with respect to a construct of interest, as is argued here, then Cell A captures what would appropriately be labeled *validity*.

[2] Messick did not use the term *consequential validity* in his 1980 or 1989 writings, but the concept is widely attributed to him as derivative from those influential works.

Although Messick did not describe it explicitly as such, the evidential basis row of the matrix (Cells A and B) can be thought of as dealing with sources of evidence that are relevant to two distinct questions: What evidence exists that the intended meaning of a test score is warranted? and What evidence exists supporting the relevance or utility of a specific test use? Put differently, the evidential basis row of the matrix can be seen as addressing the sources of evidence that are brought to bear for the two distinctly different purposes of (a) validating a test score interpretation (Cell A) and (b) justifying a specific test use (Cell B). In the consequential basis row, the value implications inherent in test score interpretation and the social consequences of test score use are highlighted.

The problems inherent in this configuration include

- isolation of value considerations (in Cell C). As Messick (1975) and others have noted, value considerations suffuse each of the enterprises represented by the four cells;
- diffusion of construct validity evidence across Cells A and B. This configuration conflates the different evidentiary requirements for test score interpretation and justifiable test use;
- distinctions among the aspects represented by each cell that are not crisply delineated; and, perhaps of greatest concern,
- inclusion of each of the distinct cellular concerns under the single heading of validity, resulting in a diffused concept—*validity*—deemed to be the most fundamental concern in test development, evaluation, and use.

Greater differentiation between the important concerns represented by each cell is clearly desirable. Messick himself noted that some important distinctions represented in the matrix "may seem fuzzy because they [i.e., the dimensions of the matrix] are not only interlinked by overlapping"; 1989, p. 20). Although Messick intended that the matrix would reduce some of the fuzziness of the distinctions, he conceded that some difficulties remained. For example, with respect to the matrix location of social consequences, Messick noted that "some measurement specialists maintain that test interpretation and test use are distinct issues, so that the two columns of [the matrix] should be treated more independently" (p. 21). Turning to the other dimension of the matrix, Messick noted that some measurement specialists "contend that value implications and social consequences of testing are separate from test validity and, as such, the two rows [of the matrix] should also be treated independently" (p. 21).

Adding to this substantial concern is the more minor problem of variability in language used to refer to portions of the matrix (e.g., bases, aspects, facets, cells, intersections). The misunderstandings and confusions about modern validity theory have been noted by many theorists and practitioners (see, e.g., Brennan, 2006; Frisbie, 2005; Hubley & Zumbo, 2011; Shepard, 1993). In blunt terms, Shepard has suggested that "the matrix was a mistake" (1997, p. 6).

Further, although each facet in the matrix addresses an important testing-related concern, those concerns are at best awkwardly fit somewhat into the single matrix intended to capture them. This problem—the lack of meaningful circumscription—has long been the subject of critique (see, e.g., Pace, 1972). Speaking to the failure of modern validity theory to delineate the boundaries of the concept, Borsboom et al. have observed that

> validity theory has gradually come to treat every important test-related issue as relevant to the validity concept. . . . In doing so, however, the theory fails to serve either the theoretically oriented psychologist or the practically inclined tester. . . . A theory of validity that leaves one with the feeling that every single concern about psychological testing is relevant, important, and should be addressed in psychological testing cannot offer a sense of direction. (2004, p. 1061)

A comprehensive treatment of those and other difficulties has been presented elsewhere (see Cizek, 2011). However, the most fundamental and consequential of the difficulties is that the distinct concerns represented in the matrix cannot all be subsumed under the single heading of validity. The essence of the problem is that of the very definition of *validity*.

The Fundamental Problem

This problem of definition is fundamental, and redressing it within the current validity framework is not possible. The solution cannot be found in reinterpretation of the matrix, by elaboration on the meaning of each cell, or by simply adding or deleting cells. An alternative framework for defensible testing is needed that treats the two fundamental testing concerns captured by the matrix—that is, validity of score inferences and justification of test use—as distinct and equally important concerns in their own right. An alternative framework to accomplish that goal will be presented subsequently. A first step in advancing validity theory and a bridge to an alternative framework is defining the key concepts *validity* and *validation.*

Perhaps the most familiar current definition of *validity* is Messick's widely cited 1989 description, which contains two components. That description defines *validity* as the act of developing "an interpretive inference . . . to ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported." However, that description continues, asserting that "to validate an inference requires validation not only of score meaning but also of value implications and action outcomes . . . and of the social consequences of using score for applied decision making" (p. 13). The same double-barreled definition can be seen in Messick's description of validation as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of *inferences* and *actions*" (p. 13, emphasis in original). In short, the contemporary conceptualization suffers from defining a purportedly unitary concept—*validity*—as two things: (a) the extent to which evidence supports an intended test score inference and (b) the extent to which the subsequent actions or consequences of using a test align with (implicit or explicit) values and intended actions or outcomes.

The problem with *validity* being currently defined as two things is compounded by the fact that it is defined as two qualitatively different things, where one aspect (i.e., accuracy of score inferences) is incompatible with the other (i.e., the justification for actions taken based on test results). The incompatibility can be illustrated with an example. Consider a hypothetical, highly accu-

rate blood test. The test reveals the presence of a certain marker that portends the onset of an incurable condition that results in a certain and agonizing death. As use of the test grows, researchers observe a rise in the incidence of suicide for persons who are informed that the test has yielded a positive result. Two things are clear from this scenario. First, it would seem legitimate to question if the test should be used at all—and if it is used, how should it be used. Further, it seems clear that ethical issues should be considered in justifying options for test use, which would include (a) discontinuing the use of the test, (b) using the test but not reporting the results to the patients, and (c) requiring that reporting of test results to patients be accompanied by counseling and follow-up. Second, the example illustrates that the meaning, interpretation, or inference of the test result—that is, the validity of the test result—is unaffected by actions based on the test scores or uses of the test results. Third, although decisions about how to use the test—or even whether to use it at all—must be made, those decisions necessarily presume that the test results are valid. However, even substantial evidence regarding score validity does not dictate what decisions about test use should be.

In the third edition of *Educational Measurement,* the term *validity* is classically defined 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" (Messick, 1989, p. 13). However, this perspective on validity requires integration of that which cannot be combined to yield a coherent result. As currently conceptualized, the process requires the gathering and synthesis of empirical evidence, theoretical rationales, and information about two important but incompatible dimensions: evidence bearing on an intended inference and evidence regarding test use. The qualitatively different sources of evidence involved in validating score meaning and justifying test use preclude making the kind of integrated, evaluative judgments demanded. An evaluative judgment based on integration of, for example, evidence based on test content or relationships among relevant variables and information about the social consequences of test use is not possible. If empirical relationships among variables failed to support—or even refuted—an intended inference, but information from test use revealed highly desirable social benefits, what integrated conclusion about validity could possibly be reached? These sources of information—evidence about the extent to which a test yields accurate inferences about a construct and evidence about the broader consequences of using the test—are not compensatory in any logical sense and cannot be combined into a coherent, integrated evaluation. Separate conclusions (i.e., conclusions about the meaning of scores and about the justification of using a test in a certain way) might be reached, but any attempted integration of the evidence confounds conclusions about both score interpretation and the desirability of using the test.

In summary, those engaged in applied testing are faced with the impossible task of producing an integrated evaluative judgment and a synthetic evaluation of that which cannot be integrated or synthesized. It is not surprising that, in over 20 years, no example of Messick's proposed outcome (i.e., a synthesis of theoretical, empirical, and social consequence data yielding an overall judgment about validity) has been produced.

Defining Validity

Thus, a starting point for a comprehensive reconceptualization that addresses both validation of score inferences and justification of test use is to first address the lack of precise definitions for the key concepts of *validity* and *validation.* The definitions that follow begin with two weak assumptions: (a) that all tests are developed with the common purpose that scores on the test will reflect variation in whatever characteristic the test is intended to measure and (b) that all tests are developed for use in at least one specific population or context. These presuppositions compel reconsideration of the first tenet of the validity canon described earlier; namely, that validity is a property of scores yielded by an instrument and not a characteristic of the instrument itself.[3] As it turns out, validity is necessarily an interaction between those two positions: It is a property of inferences about scores, generated by a specific instrument, administered under acceptable conditions to a sample of examinees from the intended population.

A straightforward definition results: *Validity is the degree to which scores on an appropriately administered instrument support inferences about variation in the characteristic that the instrument was developed to measure*. According to this definition, a test does not *define* the construct it seeks to measure but is, ideally, highly responsive to it. The situation is analogous that of a tuning fork with a specific natural frequency. Unperturbed, the tuning fork remains silent. Similarly, in the presence of a stimulus frequency other than its own, the fork does not respond. However, when a source producing the same natural frequency is placed near the tuning fork, the fork resonates. Now, the tuning fork is not the frequency; it merely responds to the presence of the frequency it was designed to produce. In like manner, a test is not the construct, and an instrument does not define a characteristic. However, a good test (i.e., one that yields accurate inferences about variation in a characteristic) will resonate in the presence of the construct it was designed to measure.

It should also be emphasized that this attribute of responsiveness does not mean that mere covariation is sufficient evidence of validity. High fidelity between variation in scores on an instrument and underlying construct variation affords some confidence that the intended inferences about the construct are appropriate. Low fidelity suggests that the characteristic being measured has not been adequately conceptualized (i.e., construct underrepresentation), that factors other than—or in addition to—the intended construct play a large role in score variation (i.e., construct-irrelevant variance), and that caution is warranted when making inferences from scores about persons' standing on the characteristic.

These principles lead to a revised definition of *validation*: *Validation is the ongoing process of gathering, summarizing, and evaluating relevant evidence concerning the degree to which that evidence supports the intended meaning of scores yielded by an*

[3] Borsboom made a similar distinction, suggesting that "a test is valid for measuring an attribute if variation in the attribute causes variation in the test scores." However, in making that distinction, Borsboom argued that "validity is a property of tests" (2005, pp. 162–163)—a position that differs from the clarification presented subsequently in this article, in which *validity* is defined not strictly as a characteristic of a test or as a characteristic of a score but necessarily as an interaction of the two.

instrument and inferences about standing on the characteristic it was designed to measure. That is, validation efforts amass and synthesize evidence for the purpose of articulating the degree of confidence that is warranted concerning intended inferences. This definition is consistent with the views of Messick (1989), Kane (1992), and others who have suggested that validation efforts are integrative, subjective, and can be based on different sources of evidence such as theory, logical argument, and empirical evidence.

An extension of the definition of *validation* forces reconsideration of a second tenet of the validity canon; namely, that all validity is construct validity. While it may be tolerable shorthand to speak of all validity as construct validity, that construction is too simplistic. It is more accurate to say that all validation is conducted for the purpose of investigating and arriving at judgments about the extent to which scores on an instrument support inferences with respect to the construct of interest.

With greater clarity about the concept of validity and the process of validation, it is now possible to distinguish the process of gathering evidence in support of intended score inferences from the process of gathering evidence to support intended test uses. In the following section, a revised heuristic is presented in which issues related to support for intended score inferences are presented as parallel to, but distinct from, issues related to test use.

Distinguishing Aims: Validation of Score Inferences and Justification of Test Use

In 1975, Messick suggested that there were two searching questions that must be asked about a test:

> First, is the test any good as a measure of the characteristic it is intended to assess? Second, should the test be used for the proposed purpose? The first question is a technical and scientific one and may be answered by appraising evidence bearing on the test's psychometric properties, especially construct validity. The second question is an ethical one, and its answer requires an evaluation of the potential consequences of the testing in terms of social values. (p. 962)

In the 2 × 2 matrix described previously, the technical and scientific question became awkwardly intertwined with the ethical and social values question. In addition to the matrix, however, Messick presented another illustration in his influential chapter to represent what he termed the "feedback representation of unified test validity" (1989, p. 90). Although it has received considerably less attention than the matrix, the figure presented there shares the same underlying assumptions as does the matrix, but it also serves as a starting point for illustrating the related but distinguishable processes of validation of score inferences and justification of test use. An alternative to that representation is shown in Figure 1. As can be seen, the figure not only incorporates the parallel priorities of validation of intended score inference (the left half of the figure) and justification of test use (the right half) but also clearly differentiates between them and highlights the inherent interactions.

As illustrated in Figure 1, the testing process begins with the necessary prerequisite of a clear statement of the intended inferential claim. This statement guides the validation effort and gathering of evidence that is then evaluated with respect to the support provided for the claim. The bidirectional arrow between the intended inferential claim and validation of score inference reflects that frequently encountered, recursive process in which the gathering and evaluation of validity evidence prompts reexamination and refinement of the intended inferential claim, which in turn suggests alternative validation strategies and sources of evidence.

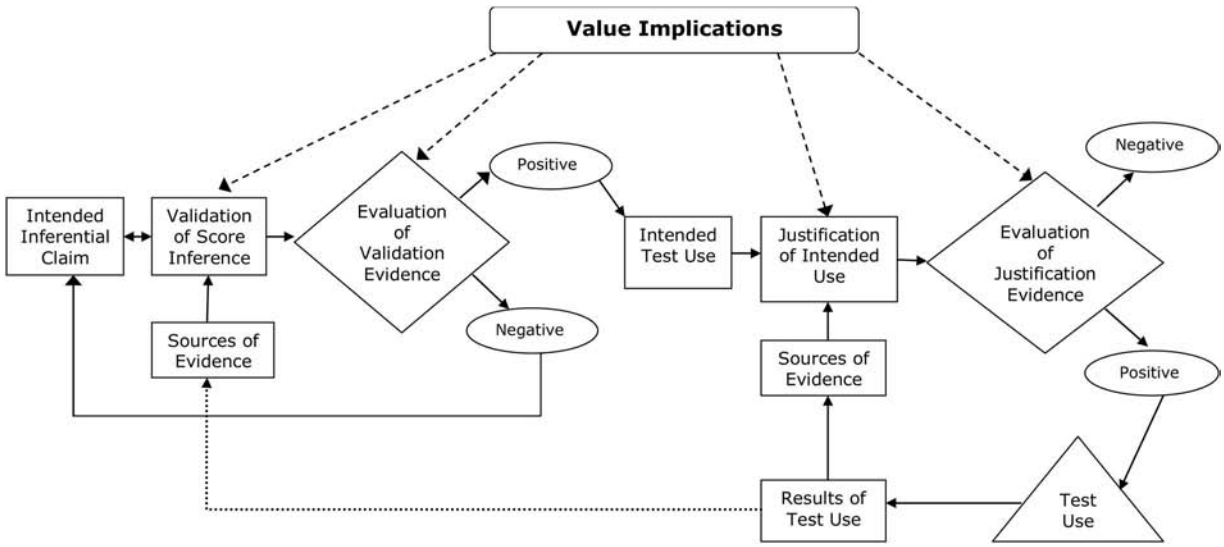

Figure 1. Relationships between validation of score inference and justification of test use. The solid lines and arrows in the figure represent a linear flow of activities (single-ended arrows) or a recursive process (double-ended arrow). The value considerations (indicated by dashed lines in the upper half of the figure) are not a similarly linear flow; rather, they permeate all of the score inference validation and score use justification process. The solid line from the Results of Test Use box indicates that results directly provide a source of evidence contributing to the corpus of justification evidence, whereas the dotted line from that box indicates that the same results might also produce evidence bearing on the intended score interpretation.

An integrated evaluation of the validity evidence then results in an overall judgment of the extent to which the evidence supports the claim (positive) or is disconfirming (negative). Until this decision point in Figure 1, the focus of investigation, evidence gathering, and evaluation has been on the validation of intended score meaning; that is, on validity. After this point, and assuming an integrated evaluation suggests that the weight of evidence supports the intended score meaning, the effort then shifts toward investigating, gathering evidence, and evaluating the justification for use of the measure. That is, validation of an intended score inference is a necessary but not sufficient condition for test use. If the body of validity evidence suggests that scores yielded by the measure can be interpreted with the intended meaning, the intended specific use of the test scores must then be justified. If more than one use is contemplated, a justification effort for each use would be required.

As with the validation process, the justification process begins with explicit articulation of the intended use, and justification relies on the gathering and evaluation of various sources of evidence. The evaluation of justification evidence also results in a positive or negative overall decision regarding the proposed use. If negative, the use of the measure for the stated purpose is rejected or a more limited or different purpose might then be contemplated. If positive, and if the test is used for that purpose, additional evidence from the test results that bears on the justification is then possible.

Although validation of an intended score inference and justification of an intended test use are different aims, they are also often inherently related. For example, as noted previously, evidence from test use can, in some circumstances, cycle back to inform the original validation effort and evaluation. How might such cycling back occur? Information gathered after a test has been administered can be mined for evidence that the construct the test purports to measure may have been inadequately specified. The two mechanisms by which this occurs are measurement errors of omission and commission. The former, *construct misspecification*, occurs when important aspects of the construct affecting performance are not included in the measurement process; the latter, *construct-irrelevant variation*, occurs when additional characteristics beyond those intended to be measured affect performance. When identified, information about construct misspecification or construct-irrelevant variation can be used to refine theory about the construct and improve the measurement instruments designed to tap it. As illustrated in Figure 1, a byproduct of the positive decision to use a test is a body of results from that use. Those results directly provide a source of evidence that contributes to the corpus of justification evidence (the solid line from "Results of Test Use" in Figure 1); however, the same results of test use might also produce evidence bearing on the intended score interpretation (the dotted line from "Results of Test Use" in Figure 1).

Finally, the figure also illustrates the key notion that value implications suffuse the entire validation and justification efforts.

Dimensions of Score Validity and Justification of Test Use

The reconceptualization of defensible testing into the dual, distinct emphases of validation and justification illustrated in Figure 1 recognizes the inherent and interconnected relationships between test score interpretation and test use. The defensibility of test use surely depends on the validity of score interpretations; just as surely, value implications infuse all aspects of test development, validation, and practice. However, it is important that concerns about validation of score inference and justification of test use be understood and pursued in their own rights. It is possible—indeed, essential—to view validity of test score interpretations and justification of test use as separable.

Along those lines, Table 1 shows several dimensions on which the efforts can be distinguished. Fuller attention to the elements captured in the table is provided shortly. For now, however, it suffices to note that the table illustrates a crisp conceptual distinction between two fundamental concerns—the validity of a test score inference and the justification for a test's use. Much of the existing validity literature deals primarily with the former, with the latter receiving attention primarily under the umbrella of what has been called consequential validity.

The distinction between validation of score inferences and justification of test use also highlights that discrepant outcomes from those investigations are entirely possible. For example, a test may be judged as useful in a case where the evidence only weakly supports the intended inference—indeed, the rationale for using a test may be that even a test with weak evidence minimally improves decisions or that its use produces an ancillary desirable outcome. Examples of this abound, from political "push surveys" that are only nominally intended to gauge candidate preference but have the effect of sensitizing voters to candidates or issues, to the increased use of constructed-response formats in achievement testing, which has been found to stimulate desirable instructional practices.[4]

Finally, it should be emphasized that the concerns of validation and justification often interact. Evidence of the validity of score inferences is a necessary but insufficient condition for recommending or sustaining a justification for test use. Validity evidence is an essential part and precursor of the justification for the use of a test—but only a part—and one that may carry greater or lesser weight in deliberations concerning test use. As Borsboom and Mellenbergh have stated in the context of tests used for selection, placement, or with the goal of bringing about changes in society at large, "validity may play an important role in these processes, but it cannot by itself justify them" (2007, p. 109).

The two fundamental concerns of validation and justification can also be viewed from the perspective of seven dimensions on which they differ. The first column of Table 1 lists these dimensions. The second column, labeled "Validity of Score Inference," outlines how each dimension applies to validation of the inference. The third column, "Justification of Test Use," provides corre-

[4] Another example is the Myers–Briggs Type Indicator (MBTI; Briggs, Myers, McCaulley, Quenk, & Hammer, 1998), which is widely used and perceived to be beneficial. Reviews of the MBTI have concluded that it lacks validity evidence but have also commended the utility of the test (see Fleenor, 2001; Mastrangelo, 2001; Pittenger, 1993). For example, Fleenor (2001) concluded that "the MBTI appears to have some value for increasing self-insight, and for helping people to understand individual differences in personality type" (p. 818). Mastrangelo (2001) cleverly captured the irony of weak empirical validity evidence and strong utility, noting that "the MBTI should not be ignored by scientists or embraced by practitioners to the extent that it currently is" (p. 819).

Table 1
Dimensions of Score Validity and Justification for Test Use

Dimension	Validity of score inference	Justification of test use
Rationale	Support for intended score meaning or interpretation	Support for specific implementation or use
Inquiry	Antecedent; primarily prior to test availability and use	Subsequent; primarily after test is made available and put into use
Focus	Primarily evidence-centered	Primarily values-centered
Tradition	Primarily psychometric and argument-based	Potentially program-evaluation- and argument-based
Warrants	Primarily technical and scientific	Primarily ethical, social, economic, political, or rhetorical
Temporal	Typically ongoing investigation to support substantive claims	Potentially recurring negotiated decision-making process
Responsibility	Primarily test developer	Primarily test user or policy maker

sponding descriptions related to the justification of test use. The presentation in Table 1 of the two dimensions, validation and justification, is purposefully parallel, intended to convey that equal priority is given to the distinct and equally important tasks of gathering evidence in support of an intended score interpretation and gathering evidence in support of justifying a particular test use.

Differentiating between validation of score inferences and justification of test use is aided by examining them vis-à-vis common dimensions on which the emphasis of each effort differs. Seven such dimensions are listed in Table 1, and the following paragraphs elaborate on each of the dimensions and on how they relate to the validation and justification efforts. It should again be noted that, although the following paragraphs highlight how validation and justification differ on the listed dimensions, the differences apply in general and are not necessarily universal. That is, there are not bright line boundaries between the validation and justification efforts; specific instances will be observed in which the relative orientations on the dimensions will be reversed. Neither are there clean bifurcations in the roles of test developers and test users with respect to validation and justification. In general, the elaborations related to the dimensions presented in Table 1 apportion greater responsibility for validation to test makers and greater responsibility for justification to test users and policy makers. However, it seems essential (or at least desirable) that meaningful collaboration involving all parties should occur when either effort is undertaken. For example, those involved in test development and validation should keep potential consequences in mind as test goals, formats, reporting procedures, audiences, intended inferences, and so forth are determined. Those involved in gathering and evaluating information to support test use will also likely rely on information gathered in validation efforts about the meaning of the test scores.

The first dimension on which validation and justification differ is the *rationale* for gathering information. As regards validity, the rationale for a validation effort is to gather support for the specific score inference. Ordinarily, the burden to plan, gather, document, and disseminate this information falls on the developer of the test. However, the burden of gathering information to support test score meaning may at times fall on a test user when, for example, there is doubt as to whether the available evidence pertains to a specific, local context or when the user contemplates an application of the test not anticipated by the test developer or supported by the available evidence. On the other hand, the rationale for information gathering for justification is to support the intended use. The justification effort may be conducted to examine the role of the test with respect to a policy decision that incorporates use of the test, to ascertain the extent to which anticipated benefits or costs are realized, or to investigate intended or unintended consequences of testing.

The second dimension on which validation and justification differ is the timing of the *inquiry*. As regards gathering of evidence to support an intended score inference, this effort is primarily antecedent to test use. Although validation is a continuing effort and evidence continues to be gathered following test use, a substantial and adequate portion of this work must occur before the operational use of a test; that is, prior to using the test in consequential ways. For commercially developed instruments, much of the effort would occur before a test was made available.

As regards justifying a test use, some empirical evidence may be available and some arguments advanced prior to test use. However, the greater part of the effort occurs primarily following the availability and operational use of a test. When a test is used primarily because of anticipated benefits, a good deal of the evidence justifying the use of a test cannot be evaluated until after the anticipated benefits would be expected to materialize. However, because the contexts of test use evolve—and the justification for a test's use in one time, setting, and population does not automatically generalize to others—the justification effort, like the validation effort, is also an ongoing endeavor.

A third dimension on which validation and justification differ is the *focus* of the effort. Although both endeavors involve information gathering, validation is primarily data-driven and evidence-centered, whereas justification examines application and highlights differences in values brought to bear in the decision-making process. A theory/application dichotomization for this dimension is an oversimplification, but the distinction is useful. On the one hand, validation efforts are grounded in the desire to operationalize a specific theoretical orientation toward a construct, to deepen understanding about a characteristic, or to aid in refining the meaning of a particular construct. It is in this sense that the aphorism "all validity is construct validity" is meaningful. On the other hand, justification efforts are—or at least can be—apathetic to whether an instrument advances basic knowledge in a discipline, extends theory, or fosters understanding of a particular construct. Justification efforts seek primarily to determine if a particular application yields anticipated benefits or promotes an outcome deemed to be desirable apart from any theory-building benefits.

The importance of the explicit consideration of these aspects cannot be overstated; they require the articulation of preferences and values. It is true that values necessarily underlie the entirety of the testing enterprise and that value considerations are equally present in both validation and justification. In the psychometric processes that guide test development and validation of score

inferences, the values that guide the enterprise may not often be explicitly articulated. They are likely to be based on implicit beliefs and assumptions (Longino, 1990), but they are nonetheless present from the outset. For example, as Messick has observed: "Value considerations impinge on measurement in a variety of ways, but especially when a choice is made in a particular instance to measure some things and not others" (1975, p. 962). Both implicit and explicit beliefs and assumptions are brought to bear in the validation effort in decisions about what sources of validity evidence are relevant, in decisions to gather some sources of validity evidence versus others, and in the weight given to some sources of evidence versus others when the evidence is summarized. As Messick (1975, citing Kaplan, 1964) has noted: "Values influence not only our choice of measure or choice of the problem, but also our choice of method and data and analysis. Values pervade not only our decisions as to where to look, but also our conclusions as to what we have seen" (p. 963).

Whereas value considerations are present and often implicit in validation efforts, they are not only present but typically more visible in the justification effort. For example, suppose it was decided to use one test instead of another based on psychometric considerations, perhaps because the predictive validity coefficient of one test exceeded that of the other. Although it may not arise as such or at all, value considerations are present in the valuing of this psychometric evidence over other possible evidence that could be used to inform the decision. However, justification deliberations are rarely conducted solely or even primarily on psychometric grounds. In contrast to validation efforts, value considerations are often the object of the justification effort. The grist of the justification mill includes aspects of testing such as feasibility, cost, time, intrusiveness, the perceived seriousness of false positive and false negative decisions, and myriad other considerations that could influence the decision to actually use a test.

Finally, deliberations about these considerations in the justification effort are important regardless of how faithfully the instrument reflects variation in a characteristic. Put differently, information gathered for validation and justification is noncompensatory. That is, a body of evidence obtained in the course of the validation effort—no matter how strong and complete—never justifies a specific use. Decisions about test use presuppose the adequacy of the validity evidence; however, cost–benefit, cost–efficiency, resource allocation priorities, feasibility, and other factors are considerations that the test user alone is in a position to evaluate.

Flowing from distinctions in focus are the fourth and fifth dimensions on which validity and justification differ: the *traditions* brought to bear and the *warrants* for conclusions. On the one hand, the validation process invokes primarily psychometric traditions for evidence gathering and interpretation; the warrants for summary evaluative judgments about the adequacy, synthesis, and interpretation of the validity evidence are primarily technical. These traditions and warrants are often—tacitly or explicitly—endorsed with near unanimity by a disciplinary community (see, e.g., Kuhn, 1962; Longino, 2002). For example, the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) represents the endorsed position of dozens of organizations regarding the sources and standards of evidence for the validation effort. Psychometric traditions to guide validation have been developed, formalized, documented, and disseminated for nearly the last 50 years.

Similarly accepted and long-standing traditions and warrants for justification of test use do not yet exist, but some likely sources of guidance are available. One rich possibility for such a tradition is that of program evaluation (see, e.g., Donaldson, Christie, & Mark, 2009; Patton, 2008; Shadish, Cook, & Leviton, 1991; Stake, 2004). The field of program evaluation provides systematic strategies for investigating the kinds of issues that often underlie justification of test use, including needs assessment, cost–benefit approaches, cost–utility analyses, and cost–effectiveness investigations. Importantly, the field of program evaluation has long recognized and incorporated the realities of differential negotiation acumen and rhetorical skill among stakeholders, the interplay of politics in evaluation decision making, and contested notions of desirable outcomes. In addition, the concepts related to fairness in testing explicated by Camilli (2006) are relevant here. The warrants related to decisions about test use are often primarily appeals to potential or real social, economic, or other benefits and ethical concerns. On such issues, broad consensus may not exist regarding the sources of evidence justifying a particular use or the weight that any particular evidence should be afforded. Standards for judging the evidence may be themselves disputed or nonexistent.

On both sides of these dimensions, similarities and disparities exist. For example, regarding the validation effort, researchers differ in the resources they bring to investigations of support for an intended inference. These disparities can result in a stronger or weaker support for an intended score inference that may be unrelated to the quality of an instrument. Regarding justification of test use, disparities also exist; they may include resource differentials but also involve differentials in the power, position, rhetorical skill, and other characteristics of those involved in—or excluded from—the negotiation and evaluation process. Indeed, these disparities can function as gatekeepers in determining what information related to test use is brought to bear. Disparities in power and position that are manifested in the course of justifying test use do not affect support for the intended inference (i.e., the validity argument); they can, however, serve to frame arguments about legitimate or inappropriate test uses.

The sixth dimension that distinguishes the validation and justification efforts is *temporal*. As indicated previously, a tenet of modern validity theory is that the case for validity is never truly considered to be closed. Efforts aimed at gathering evidence to support intended test score inferences should be ongoing; judgments about the appropriateness and accuracy of the intended inference are continually buttressed or threatened by the accumulation and analysis of new information, diverse applications, additional research, theoretical refinements, and other factors that yield evidence after the initial validation effort.

In contrast, although information gathered in justification of a test use should also be gathered continuously, the effort is comparatively more decision-oriented. The decision to use a test for a specific purpose is a process that incorporates the input and perspectives of those who allocate resources, policy makers, constituencies, stakeholders, and others affected by the use of a test. The process has as its immediate aim the practical outcome of deciding if and how a test should be used in a specific way. Thus, information gathering in support of test justification is more goal-oriented and focused on the decision at hand than the information gathering of a validation effort. The temporal aspect of justification of test use is also apparent in that the case for a specific use

may be deemed weaker or stronger or demand reconsideration to the extent that there is a change in context, policy, political goals, or stakeholders or a shift in power relations. Reconsideration of test use may not be undertaken under stable conditions but may be demanded with changes in resources, when alternative procedures become available, when different policy aims prevail, when unintended consequences are discovered, or when questions arise about whether test results are being used in a manner that meets a perceived need or goal.

The final dimension on which validity of score inferences and justification of test use differ concerns the ultimate *responsibility* for the effort. Because evidence of validity should be developed prior to making a test available, it is the test developer who has the potential to engage in this analysis and who bears singular responsibility for establishing the validity of intended score inferences. On this count, the framework presented here is especially powerful toward the goal of improving the practice of validation. The process for articulating a series of related, logical claims about the score inference—what Kane (1992, 2006, 2009) has described as an argument-based approach—provides a straightforward starting point for contemplating the appropriate sources of evidence and judging whether the claims are adequately supported.

As regards justification of test use, the responsibility for deciding upon appropriate sources of evidence to justify a particular use; for gathering that evidence; and for developing an integrated, evaluative judgment concerning test use is likely to be apportioned in different ways. In some cases, a test developer will not only have in mind a particular score interpretation but might also intend a specific use for which the instrument is to be commended or marketed. In such cases, both the validation and justification burdens would fall on the test developer. In other cases, a test might be considered for a use that the test developer never envisioned. In such cases, the responsibility for justification of test use would fall squarely on the user. Even in situations where a test use might reasonably be anticipated by a test developer, the burden of justifying a specific test use would seem to fall more directly on the test user. It is the test user who decides to employ a test in the first place; it is the test user who, when options exist, chooses one test over another; and it is the test user who typically associates consequences, rewards, sanctions, or decisions with test performance. As Kane has indicated, "an argument can be made for concluding that the decision makers (i.e., the test users) have the final responsibility for their decisions . . . and they are usually in the best position to evaluate the likely consequences in their contexts of the decisions being made" (2001, p. 338). Finally, in still other cases, a test developer may not have any specific interest (except, perhaps, a pecuniary one) in a test use and would not ordinarily be in a position to aid decision makers with the policy calculus they must perform. Nonetheless, because of the perspective of the test developer and the insights gained in the validation effort, a collaborative justification effort between test developer and test user would be seem advantageous.

Critical Commonalities

The seven dimensions just described reveal that validation of score inferences can—and should—be distinguished from justification of test use. Although the presentation has highlighted differences between the validation and justification efforts, they also share common characteristics.

First, both the validation and justification efforts require an evaluative integration of information to arrive at a conclusion: One conclusion is a judgment about how adequately the evidence supports the intended score inference; the other is a judgment about how compelling the case is for a specific application.

Second, the constellation of evidence brought to bear for either effort is necessarily dependent upon the particular validation or justification situation at hand. The essential sources of evidence to support the validity of a score inference and the sources of evidence to justify a test use vary across test purposes, contexts, policies, and risks; the mix of evidence that is appropriate for one situation may be inappropriate for another. The determination of appropriate sources of validation evidence depends on the specific inference(s) that scores are intended to yield; the determination of appropriate sources of justification evidence depend on the specific use to which the scores will be put.

For example, imagine a validation effort for a hypothetical test claimed to support inferences about a multifaceted construct, with subscores formed on the basis of constellations of items purporting to measure different aspects of the construct. At minimum, the validation effort would seem to require empirical evidence about the internal structure of the instrument. Factorial validity evidence would buttress claims about the intended inferences; an absence of such evidence would pose a serious threat to claims about what the instrument measures. Of course, factorial validity evidence alone would be insufficient because variation in test performance may still be attributable to something other than the construct of interest; other sources of evidence would need to be mined, based on the specific claims (see Kane, 2006). For example, documentation that the items were created and that the constellations of items were formed to reflect the theory underlying the construct (i.e., evidence based on test content; evidence based on test development process), as well as evidence that scores on the instrument do not reflect covariation with another, different construct (i.e., evidence based on relationships with other variables), would be necessary.

In assembling the case to justify a specific use of this hypothetical test, the appropriate evidentiary sources would depend on the particular context, intended outcomes, values, and constituencies involved in the evaluation. For example, suppose that the test was determined to have adequate validity evidence to support the intended inferences based on its subscores. Justification evidence would now be required, including information such as the personnel time required to appropriately administer and interpret the test results, alternative measures that could be used, consideration of other priorities for spending the resources allocated to testing, or information about the usefulness of the subscore information to clinicians or the test takers themselves. In all cases, the sources of evidence for justification of test use would be grounded in the priorities, policies, power, and ethical considerations of the parties involved in the decision-making process.

Third, the question of "How much evidence is sufficient?" represents a continuing dilemma in validation; the question also pertains to justification. Whether for validation or justification, marshaling as much expertise and information as possible will continue to provide the best foundation for decisions about inference or use. As Kane has indicated: "different interpretations/uses will require different kinds of

and different amounts of evidence" (2009, p. 40). Judgments about the appropriate sources and quantity of evidence for validation efforts depend on the nature, breadth, and complexity of the intended inference; the relative seriousness of inaccurate inferences; and the resources available for the validation effort. Judgments about the appropriate sources and quantity of evidence for justification of test use depend on the nature, breadth, and complexity of the intended use; the personal, systemic, and social consequences of that use; and the resources available for the justification effort. Ultimately, because both validation and justification efforts are grounded in the priorities, policies, power, resources, and ethical considerations of those involved—and because both require the application of expertise, evaluation of evidence, and values—it is unlikely that a standard could be established to specify the sources or extent of evidence that are "enough."

Implications and Conclusion

Following from a revised framework in which validation of score inferences and justification of test use are considered as separate, parallel, and equally valued endeavors are at least four immediate implications. First, by providing clarity regarding the definition of *validity*, the proposed reconfiguration suggests answers to some lingering issues. For example, a question that has not been satisfactorily answered to date centers on whether the validity of test scores depends on their use. The differentiation between validation of score inferences and justification of test use clarifies that issue by disentangling what is revealed to be, in essence, two questions. Within the proposed reconceptualization, the intended meaning of the score with respect to the measured construct (i.e., *validity*) is established during the test development and validation efforts. That is, the validity or meaning of the score with respect to the intended construct does not depend on the specific use of the test. The usefulness of the score does depend, however, on the various contexts, decisions, or situations in which the test is applied. This is a separate effort in which empirical evidence and logical rationales must be conducted to determine if the (validated) score meaning is relevant to and justifiably used in the service of the diverse applications to which the test will be put.

Second, the proposed reconceptualization provides a home for important concerns that heretofore have not readily fit into the existing validity framework. For example, under the umbrella "Justification of Test Use," the proposed framework clarifies and provides an obvious home for what has been referred to as consequential validity—the concern about how best to account for the personal, systemic, social, and other consequences of testing. Other examples that are brought under the same umbrella are the concepts of *fairness* and *opportunity to learn.* Fairness—which at first blush is sometimes cast as a validity concern—is actually treated as a separate, independent chapter in the most recent edition of *Educational Measurement* (see Camilli, 2006). Questions about fairness, adverse impact, differential prediction, and related concerns are clearly subsumed under and relevant to justification of the use of a test. Likewise, as a characteristic of the defensible use of educational achievement tests since at least the 1970s, opportunity to learn has languished without a clear place in validity theory. Under the proposed reconfiguration, that concept also fits squarely as a relevant source of evidence when building the case that the use of a test is justified.

A third implication suggests at least some of the work that lies ahead. Whereas established standards and traditions for validation of score inferences have existed for some time, similar standards and traditions for justification of test use have not yet been developed. However, in addition to the previously mentioned and potentially helpful paradigmatic lenses that might be adapted from the practice of program evaluation, at least one model for developing justifications in support of a specific test use is also available. Kane's (1992, 2006, 2009) procedures for pursuing an argument-based approach to validity might be profitably employed to craft argument-based approaches to justification of test use. As Kane (1992, 2006) has indicated, the choice of the term *argument-based* was appropriate for referring to an approach, and an argument-based approach is consistent with the reconceptualization presented here. In the context of the dual investigations that must occur (i.e., validating intended score interpretations and justifying a specific test use), the conditions that Kane (1992) described for an argument-based approach to validation seem equally relevant to justification; namely, there is an audience to be persuaded, there is a need to develop a positive case for the interpretation or use, and there is a need to consider and evaluate competing conclusions.

A final implication has to do with scope of application. As Zumbo has observed, "It is rare that anyone measures for the sheer delight one experiences from the act itself. Instead, all measurement is, in essence, something you do so that you can use the outcomes" (2009, p. 66). Along these lines, the proposed framework can be seen to have broad applicability. It is equally applicable to commercial tests, published measures, surveys, scales developed for research—indeed, to any context in which an instrument is developed to tap a specific construct and when the administration and use of the instrument can be anticipated to affect the persons, systems, or contexts in which the scores will be used.

In conclusion, it is clear that validity theory has advanced appreciably, and it continues to be an evolving concept. Modern psychometrics has moved far from the notion that "a test is valid for anything with which it correlates" (Guilford, 1946, p. 429) to a more sophisticated paradigm with many broadly accepted, fundamental tenets. The framework presented here builds on those fundamentals and proposes a more comprehensive framework that subsumes concerns about score meaning and test use while differentiating these related inquiries into two parallel endeavors. The first endeavor is one that gathers and evaluates support for test score inferences; that is, *validation.* The second endeavor is one that gathers and evaluates support for test use; that is, *justification*.

Of course, reconceptualization alone will not resolve the oft-noted gap between validity theory and validation practice: Greater alacrity in validation and justification efforts is still required. To foster the latter, much work must be undertaken to ramp up research and practice with respect to justification. There is broad consensus and fairly well established standards and methods for validation of score inferences. However, little work has been done to articulate guidelines or procedures for justification of test use. One hoped-for consequence of differentiating between validity of score inferences and justification of test use is that rigor regarding both efforts will also advance and parallel each other.

To the extent that the concept of validity is more clearly focused and justification efforts are stimulated, a revised framework can help foster the goals of facilitating more complete and searching

validation practice, enhancing the quality and utility of test results, and enabling those who develop and use tests to improve the outcomes for the clients, students, organizations, and others that are the ultimate beneficiaries of high-quality test information.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *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.* Washington, DC: Author.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511490026

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton & M. J. Gierl, *Cognitive diagnostic assessment for education* (pp. 85–116). Cambridge, England; Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071. doi:10.1037/0033-295X.111.4.1061

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: Praeger.

Briggs, K. C., Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *Myers-Briggs Type Indicator—Form M.* Mountain View, CA: Consulting Psychologists Press.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: Praeger.

Cizek, G. J. (2011, April). *Error of measurement: Reconsidering validity theory and the place of consequences.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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

Donaldson, S. I., Christie, C. A., & Mark, M. M. (Eds.). (2009). *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: Sage.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16,* 640–647. doi:10.1037/h0045478

Fleenor, J. W. (2001). Review of Meyers-Briggs type indicator. In B. S. Plake & J. C. Impara (Eds.), *The fourteenth mental measurements yearbook* (pp. 816–818). Lincoln, NE: Buros Institute of Mental Measurements.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24*(3), 21–28. doi:10.1111/j.1745-3992.2005.00016.x

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist, 27,* 197–222. doi:10.1207/s15326985ep2702_5

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6,* 427–439.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11,* 385–398. doi:10.1037/0735-7028.11.3.385

Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology, 19,* 451–473. doi:10.1177/0959354309336320

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103,* 219–230. doi:10.1007/s11205-011-9843-4

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527–535. doi:10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 39–64). Charlotte, NC: Information Age.

Kaplan, A. (1964). *The conduct of inquiry.* San Francisco, CA: Chandler.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago, IL: University of Chicago Press.

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

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry.* Princeton, NJ: Princeton University Press.

Longino, H. E. (2002). *The fate of knowledge.* Princeton, NJ: Princeton University Press.

Markus, K. A. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research, 45,* 7–34. doi:10.1023/A:1006960823277

Mastrangelo, P. M. (2001). Review of Meyers-Briggs type indicator. In B. S. Plake & J. C. Impara (Eds.), *The fourteenth mental measurements yearbook* (pp. 818–819). Lincoln, NE: Buros Institute of Mental Measurements.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16–18. doi:10.1111/j.1745-3992.1997.tb00588.x

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966. doi:10.1037/0003-066X.30.10.955

Messick, S. (1980). *Test validity and the ethics of assessment American Psychologist, 35,* 1012–1027. doi:10.1037/0003-066X.35.11.1012

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

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35–44. doi:10.1023/A:1006964925094

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12. doi:10.1111/j.1745-3992.1998.tb00826.x

Pace, C. R. (1972). Review of the Comparative Guidance and Placement Program. In O. K. Buros (Ed.), *The seventh mental measurements yearbook* (pp. 1026–1028). Highland Park, NJ: Gryphon.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Newbury Park, CA: Sage.

Pittenger, D. J. (1993). The utility of the Myers-Briggs Type Indicator. *Review of Educational Research, 63,* 467–488.

Popham, W. J. (1997). Consequential validity: Right concern, wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13. doi:10.1111/j.1745-3992.1997.tb00586.x

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practices.* Newbury Park, CA: Sage.

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

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. doi:10.1111/j.1745-3992.1997.tb00585.x

Stake, R. E. (2004). *Standards-based and responsive evaluation.* Thousand Oaks, CA: Sage.

Tenopyr, M. L. (1966, April). *Construct-consequences confusion.* Paper

presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Wainer, H., & Braun, H. I. (Eds.). (1988). *Test validity*. Hillsdale, NJ: Erlbaum.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (pp. 45–79). Amsterdam, the Netherlands: Elsevier Science.

Zumbo, B. D. (2009). Validity as contextualized as pragmatic explanation and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Charlotte, NC: Information Age.

Received October 7, 2009
Revision received December 13, 2011
Accepted December 16, 2011 ■