ASSESSMENT IN ACTION

The Meaning of Validity in the New *Standards for Educational and Psychological Testing:* Implications for Measurement Courses

Laura D. Goodwin and Nancy L. Leech

The treatment of validity in the newest edition of Standards for Educational and Psychological Testing *(*Standards*; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) is quite different from coverage in earlier editions of the* Standards *and in most measurement textbooks. The view of validity in the 1999* Standards *is discussed, and suggestions for instructors of measurement courses are offered.*

In the newest edition of the *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), the treatment of validity is markedly different from what it was in the three earlier editions of the *Standards* (AERA, APA, & NCME, 1985; APA, 1954; APA, AERA, & NCME, 1966). The purpose of this article is to describe the meaning of validity in the new *Standards*, to compare that meaning with the presentation of validity in commonly used measurement textbooks, and to discuss implications for the teaching of validity in measurement courses for counseling students.

BACKGROUND

As Geisinger noted in 1992, the meaning of validity had undergone a "metamorphosis" during the previous half-century. In addition to Geisinger, other measurement experts and theorists who described historical changes in the definitions of validity include Angoff (1988), Cronbach (1988, 1989), Goodwin (1997, 2002a), Kane (1994, 2001), Messick (1988, 1989a, 1989b), Langenfeld and Crocker (1994), Moss (1992), and Shepard (1993). With the publication of the new *Standards* in 1999 (APA, AERA, & NCME, 1999), Geisinger's statement about the metamorphosis of validity definitions is as pertinent today as it was more than 10 years ago.

An early definition of validity, from the 1940s, emphasized the test itself. Validity was conceptualized as a static property of a measure—a view epitomized by Guilford's (1946) often cited statement that, "in a very general sense, a test is valid for anything with which it correlates" (p. 429). Other well-known psychometricians of the time who held the same general view of validity included Cureton (1951) and Gulliksen (1950). A test was considered to be either valid or not as evidenced by the correlations between the test and some other "external" criterion measure.

Laura D. Goodwin and Nancy L. Leech, School of Education, University of Colorado at Denver. Correspondence concerning this article should be addressed to Laura D. Goodwin, School of Education, University of Colorado at Denver, PO Box 173364, Denver, CO 80305 (e-mail: Laura.Goodwin@cudenver.edu).

With the publication of the 1966 *Standards* (APA, AERA, & NCME, 1966), the meaning of validity shifted its focus to use—that is, validity was defined as the extent to which a test produced information that was useful for a specific purpose. This edition of the *Standards* also included the "trinity" view of validity first presented by Cronbach and Meehl in 1955. Validity was categorized into specific types: content validity, criterion-related validity (which included concurrent and predictive validities), and construct validity. At about the same time, Campbell and Fiske (1959) extended the dialogue about discrete types of validity—and the need for multiple kinds of validity evidence—with their landmark presentation of the multimethod-multitrait approach to validation, which included the introduction of convergent and discriminant types of validity.

During the 1980s, another shift in the conceptualization of validity emerged. Led by Cronbach (1980, 1988) and Messick (1980, 1988, 1989b), psychometricians and measurement experts began to emphasize the inferences and decisions made from test scores. In the 1985 *Standards* (AERA, APA, & NCME, 1985), validity was defined as "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" and test validation was described as "the process of accumulating evidence to support such inferences" (p. 9).

The 1980s and 1990s witnessed two other important additions to the dialogue about the meaning of validity. The usefulness of the trinity view of validity was being challenged. Although the 1985 edition of the *Standards* continued to include this method of categorizing types of validity, the authors cautioned that, "the use of the category labels should not be taken to imply that there are distinct types of validity" (AERA, APA, & NCME, 1985, p. 9). Increasingly, validity was conceptualized as a unitary concept—with construct validity being the key and unifying type of validity (Langenfeld & Crocker, 1994; Messick, 1989b, 1995; Shepard, 1993). Among the prominent measurement theorists who wrote about the obsolete nature of the trinity view of validity was Brennan (1998), who described the content-criterion-construct breakdown as "somewhat arbitrary, certainly incomplete, and clearly not [a] coequal set of categories of evidence for validating inferences" (p. 7). Kane (1994) also wrote the following: "Although many kinds of evidence may be used, we do not have different kinds of validity. Validity involves an overall evaluation of the plausibility of the intended interpretations" (p. 136). Langenfeld and Crocker predicted that the next edition of the *Standards* would not include the traditional breakdown of validity into three parts—a prophecy that was fulfilled with the publication of the most recent *Standards* in 1999 (AERA, APA, & NCME, 1999).

Also gaining attention during the 1980s and 1990s was the need for evidence about the social *consequences* of test use (Cronbach, 1988; Linn, 1994; Messick, 1989b, 1994; Shepard, 1993). As Shepard (1993) pointed out, this issue was not entirely new: In the first edition of *Educational Measurement*, Cureton (1951) wrote that "the essential question of test validity is how well a test does the job it was employed to do" (p. 621). What was new (and controversial) in discussions about the role of consequences in validation research was studying both intended and unintended—often adverse—consequences of test use. The advantages and disadvantages of including investigations about consequences as part of validation efforts were the topics of presentations at national conferences (e.g., AERA and NCME) as well as an entire issue of *Educational Measurement: Issues and Practice* (Linn, 1997; Mehrens, 1997; Popham, 1997; Shepard, 1997). Some measurement experts (e.g., Kane, 2001; Linn, 1997; Shepard, 1997) have argued for the broader conceptualization of validity (one that includes the consequences of using tests and other measures), whereas others (e.g., Popham, 1997) have advocated for a more limited and technical definition of validity that focuses primarily on the descriptive interpretation of scores. Including consequential evidence as part of validity has been controversial and not well accepted by some because investigating consequences extends beyond traditional psychometric boundaries into policy arenas (Dwyer, 2000); it is what Kane (2001) termed the *prescriptive* part of a validity argument:

> In discussing the role of consequences in validation, it would probably be useful to separate the interpretive argument into two parts. The *descriptive* part of the argument involves a network of inferences leading from scores to descriptive statements about individuals, and the *prescriptive* part involves the making of decisions based on the descriptive statements. For example, the use of a reading comprehensive test to place students in reading groups involves conclusions about each student's level of reading skill, and then a decision about placement, which may involve additional information or constraints (e.g., group sizes). (p. 327)

Although debates about the role of consequences in validity are likely to continue for some time, the authors of the 1999 *Standards* (AERA, APA, & NCME, 1999) did include consequential evidence as a type of validity evidence for the first time. It is important to note, however, that this is not a new type of validity—and the term *consequential validity* should be avoided.

VALIDITY AS DESCRIBED IN THE NEW *STANDARDS*

In the newest edition of the *Standards*, the chapter on validity presents the following description of validity and the validation process:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated. (AERA, APA, & NCME, 1999, p. 9)

Prior to describing five distinct types of validity evidence, the authors of the *Standards* point out the difference between the types of evidence and types of validity:

> These sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all of the accumulated evidence supports the intended interpretation of test scores for the intended purposes. (AERA, APA, & NCME, 1999, p. 11)

Evidence Based on Test Content

Relevant for almost all measures and measurement procedures, this type of validity evidence is based on logical analyses and experts' evaluations of the content of the measure, including items, tasks, formats, wording, and processes required of examinees. In general, it addresses questions about the extent to which the content of a measure represents a specified content domain (Goodwin, 2002a). Experts' reviews are conducted to obtain evidence of such features of a test as sufficiency, clarity, relevancy, and the match between the items and tasks and the definition of the construct. Additional aspects of the review process often focus on bias (gender, culture, age, etc.). Included here are the notions of "construct-irrelevant variance" and "construct underrepresentation," meaning the extent to which the test seems to measure more or less than what is intended. Either problem can give unfair advantages to one or more subgroups of respondents.

Given the fact that this type of validity evidence is needed for virtually all measures, as well as its importance in court decisions resulting from challenges to the validity of certification and licensing examinations (Sireci & Green, 2000), it is curious that it receives relatively little attention in measurement textbooks (Berk, 1990)—although many textbooks do include information about using a table of specifications to link the measure's domain or universe with the specific items or tasks. Guidelines on such concerns as the optimal numbers of experts, the selection and training of the experts, and the collection

and analysis of content-related evidence have been offered by a few researchers (e.g., Berk, 1990; Grant & Davis, 1997), but more information of this type in textbooks would be a welcome addition.

Evidence Based on Response Processes

In the previous edition of the *Standards* (AERA, APA, & NCME, 1985), this type of evidence was included as an aspect of construct-related validity evidence. It examines the extent to which the tasks or types of responses required of examinees fit the intended, defined construct. For many self-report, paper-and-pencil measures of constructs belonging to the affective domain, evidence that respondents are not just giving socially desirable answers is an example of this type of evidence. Ways to obtain evidence of this sort include observing examinees as they perform required tasks and interviewing examinees to determine reasons for providing certain answers to questions. Also included here are investigations of the ways in which observers, judges, and raters use criteria to record and evaluate behaviors, performances, essays, and so forth. "The concern is that they are applying the criteria as intended and not using irrelevant or extraneous factors that do not match the planned interpretations of scores" (Goodwin, 2002a, p. 102).

Evidence Based on Internal Structure

Like the evidence based on response processes, this type of evidence also was considered part of construct-related evidence in the 1985 *Standards* (AERA, APA, & NCME, 1985). It examines the extent to which the internal components of a test match the defined construct and is most often estimated by confirmatory factor analysis. Given that factor analytical evidence is relatively easy to obtain, researchers sometimes place too heavy an emphasis on it—even going so far as to place sole emphasis on it for validation purposes (Goodwin, 1999). As other measurement experts have noted (e.g., Nunnally & Bernstein, 1994; Thorndike, 1997), placing too much reliance on factor analysis for validity evidence can result in a very narrow body of empirical support for validity arguments.

The use of differential item function (DIF) techniques to help detect item bias is also included in this category of types of validity evidence. "DIF is present when examinees of the same ability but belonging to different groups have differing probabilities of success on an item" (Hattie, Jaeger, & Bond, 1999, p. 432). Various ways of assessing DIF have been proposed, some based on classical test theory and some based on item response theory (Nunnally & Bernstein, 1994). The Mantel-Haenszel technique is one specific approach to the study of DIF (Holland & Thayer, 1988; Silverlake, 1989).

Evidence Based on Relations to Other Variables

This category of evidence is the most extensive category and encompasses many of the old specific types of validity: criterion-related validity (including concurrent and predictive validity) as well as much of what was traditionally covered under construct validity (including convergent and discriminant validity). The most commonly used approaches to the collection of this kind of validity evidence are correlational studies, criterion-group or known-group comparison studies, and experimental studies. The correlational studies include investigations of the nature and extent of relationships between scores and external, "criterion" variables (obtained simultaneously or at a later date); studies of the relationships between scores and data obtained with other instruments or measurement techniques intended to measure the same constructs; and studies of the relationships between scores and measures of purportedly different constructs. Group-comparison studies are often used in validity research aimed at testing hypotheses about expected differences in scores across various groups of examinees and differential group prediction or relationship stud-

ies. Experimental research studies also can be conducted to test hypotheses about effects of interventions on scores, as well as the effectiveness of placement, selection, and classification decisions. Finally, studies of validity generalization (Schmidt & Hunter, 1977) are part of this category of validity evidence types as well.

Evidence Based on the Consequences of Testing

This type of validity evidence pertains to anticipated and unanticipated consequences—both positive and negative—of measurement:

> Tests are commonly administered in the expectation that some benefit will be realized from the intended use of the scores. A few of the many possible benefits are selection of efficacious treatments for therapy, placement of workers in suitable jobs, prevention of unqualified individuals from entering a profession, or improvement of classroom instructional practices. A fundamental purpose of validation is to indicate whether these specific benefits are likely to be realized. (AERA, APA, & NCME, 1999, p. 16)

The consequential aspect of validity was mentioned only briefly in the 1985 *Standards* (AERA, APA, & NCME, 1985) and was not considered to be a separate type of validity. As mentioned previously, including consequential evidence as part of a comprehensive body of validity evidence is a relatively new idea and has been the subject of debate and some controversy. Because of its recent introduction into the realm of validation, few guidelines have been suggested on how to study measurement consequences (Millman, 1998). However, Chudowsky and Behuniak (1998) offered suggestions about the use of focus groups to investigate consequences of a large-scale assessment program.

IMPLICATIONS FOR MEASUREMENT COURSES

For many instructors of measurement courses, the changes in the treatment of validity in the new *Standards* are very welcome. The old tripartite view of validity, which has been included in measurement textbooks for many years, is problematic for a number of interrelated reasons. First, the categorization of validity into three (or four) types of validity is often confusing to students because of the overlap between construct validity and the other types of validity. Students easily become worried about being able to correctly identify the type of validity in a real or hypothetical example, and they lose sight of the "big picture." Instead of spending valuable class time on the important issues about validity, time is often spent on sorting out the types of validity. A second problem with the old system is that it encourages a "checklist" approach to validity when students are critiquing the extant validity evidence in test manuals or research articles. Third, the tripartite categorization scheme tends to promote the misconception that validity is a static property of an instrument or measure, rather than a property of a sample (Tracey & Glidden-Tracey, 1999). This problem leads to the inappropriate use of language about validity: for example, the use of phrases such as, "This test is valid," "This test has been proven to be valid," "This measure has demonstrated validity," and so forth. Fourth, this way of thinking about validity runs counter to notions that validity is a unitary concept (Geisinger, 1992) and that construct validity is really the "whole" of validity theory (Shepard, 1993). Finally, it too easily allows students to view all of the old types of validity as equally important in the collection of a body of evidence to support the inferences to be made from test scores.

We recently reviewed the coverage of validity in a sample of 10 measurement textbooks in order to determine the extent to which coverage of validity in those textbooks mirrors the presentation of validity in the new 1999 *Standards*. (A list of the textbooks reviewed is in the Appendix.) We chose textbooks that were written primarily for measurement courses in education (including counselor education and counseling psychology) and had fairly recent copyright dates. All of the textbooks reviewed included the tripartite categoriza-

tion of types of validity. Although the authors of several of these books mentioned the unified nature of validity, the need for multiple types of validity evidence, and the fact that the categories should not be viewed as mutually exclusive, they all nevertheless described the traditional three (or four) types of validity and devoted the majority of the coverage of validity to this topic. Although we did not conduct a systematic review of the coverage of validity in research methods textbooks, our informal observation is that most authors of those textbooks also tended to present the long-standing tripartite view of validity and provided relatively little information about other concerns, such as the meaning and importance of developing a validity argument (Cronbach, 1988) and including the study of consequences in validation efforts.

Given the publication dates of the measurement textbooks, and the lag time between revisions of textbooks and actual publications, it is understandable that the validity coverage in textbooks does not match the 1999 *Standards'* treatment of validity. (Authors of any text published prior to 2000 would not have had the opportunity to review the 1999 *Standards* and incorporate the newer material into their textbooks.) Until measurement (and research methods) textbooks are revised to reflect the presentation of validity in the new *Standards*, instructors of measurement and research methods courses will find it challenging to blend the textbook coverage of validity with the newer approach to defining and discussing validity. Furthermore, until licensing examinations (e.g., the National Counselor Examination [NCE]) change the content of the questions pertaining to measurement validity, the varying views of validity should be covered in measurement courses. The manner in which validity is typically presented in test manuals represents another challenge for the measurement community. Supplementary textbook and examination materials (instructors' guides, examination review guides, and the like) will also need to catch up with the new presentation of types of validity evidence as found in the 1999 *Standards* (AERA, APA, & NCME, 1999).

Suggestions for Instructors of Measurement Courses

For many instructors of measurement and research methods courses, incorporating the new views of validity from the 1999 *Standards* will require a shift in their own thinking about validity and in their course presentations and materials. Because many instructors learned about validity on the basis of the traditional tripartite view, such a shift could be quite challenging. Until textbooks, test manuals, and examinations such as the NCE alter the presentation of validity to better match the 1999 *Standards*, it behooves measurement course instructors to include both the traditional and the newer views of validity in their courses. The following are some suggestions for these instructors that might help with the transition.

One suggestion is to adopt the 1999 *Standards* as either a required or optional book for the course. This would allow students to read firsthand about the types of validity evidence and compare and contrast this treatment of validity with what they will find in their textbooks. In addition, organizing the in-class presentations and discussions about validity around the five types of evidence and tying this to the traditional ways of thinking about validity might help students better understand the historical shifts in the views of validity and enable them to make the transition from the tripartite view to the newer view. In Table 1, the types of validity evidence as presented in the 1999 *Standards* are listed along with the names of the corresponding types of validity from the 1985 *Standards* (where possible); also in Table 1 are examples of validation activities for each type of evidence. Table 1, adapted from Goodwin (2002a), would be a useful way to highlight the differences in the meaning of validity over time and to stimulate class discussions about types of validity evidence needed for real and hypothetical measures.

A second suggestion is to emphasize the role of arguments in validation. According to Cronbach (1988), a validity argument is a coherent analysis and evaluation of all of the evidence for and against proposed interpretations of scores including, if possible, evidence regarding plausible alternative interpretations. Haertel (1999) wrote about validation as a

TABLE 1

Types of Validity Evidence (1999 vs.1985 *Standards*) and Examples of Validation Activities

Types of Validity Evidence		
1999 *Standards*	**1985 *Standards***	**Examples of Validation Activities**
Evidence based on test content	Construct-related evidence	1. Logical analyses and experts' reviews of the extent to which the content of the measure represents the content domain or universe. 2. Logical analyses and experts' reviews of the extent to which the items, tasks, or subparts of a measure fit the definition of the construct and/or the purpose of the measure. 3. Logical analyses and experts' reviews of the relevance, importance, clarity, and lack of bias in the measure's items or tasks. 4. Logical analyses and experts' reviews of the extent to which construct underrepresentation or construct-irrelevant aspects of the measure may result in unfair advantages for one or more subgroups of respondents.
Evidence based on response processes	Construct-related evidence	1. Analyses of individuals' responses to items or tasks, via interviews with respondents. 2. Studies of the similarities and differences in responses supplied by various subgroups of respondents. 3. Studies of the ways that raters, observers, interviewers, and judges collect and interpret data. 4. Longitudinal studies of changes in responses to items or tasks.
Evidence based on internal structure	Construct-related evidence	1. Factor- and cluster-analytical studies. 2. Analyses of item interrelationships, using item analysis procedures. 3. Differential item functioning (DIF) studies.
Evidence based on relations to other variables	Criterion-related evidence (1–6); construct-related evidence (7–11)	1. Correlational studies of the strength and direction of the relationships between the measure and external "criterion" variables. 2. Correlational studies of the extent to which scores obtained with the measure predict external "criterion" variables measured at a later date. 3. Group separation studies, based on decision theory, that examine the extent to which scores obtained with an instrument accurately predict outcome variables. 4. Differential group relationship or prediction studies. 5. Studies of the effectiveness of selection, classification, and placement decisions. 6. Validity generalization studies. 7. Convergent validity studies that examine the strength and direction of the relationships between the measure and other variables that the measure should, theoretically, have high correlations with. 8. Discriminant validity studies that examine the strength and direction of the relationships between the measure and other variables that the measure should, theoretically, have low correlations with. 9. Experimental studies that test hypotheses about effects of interventions on scores obtained with an instrument. 10. Known-group comparison studies that test hypotheses about expected differences in average scores across various groups of respondents. 11. Longitudinal studies that test hypotheses about expected differences in average scores over time.
Evidence based on consequences of testing		1. Studies of the extent to which expected or anticipated benefits of testing are realized. 2. Studies of the extent to which unexpected or unanticipated negative consequences of testing occur.

Note. From "Changing Conceptions of Measurement Validity: An Update on the New *Standards*," by L. D. Goodwin, 2002, *Journal of Nursing Education, 41*, p. 103. Copyright 2002 by SLACK, Inc. Adapted with permission.

process of constructing and evaluating arguments for and against proposed test score uses and interpretations: "Any such argument will involve a number of assumptions, or propositions, each of which requires support. The overall argument is only as strong as its weakest premise" (p. 5). He continued by providing an example of posing validity arguments for large-scale testing programs.

Tracey and Glidden-Tracey (1999) offered an interesting set of ideas regarding construct specification. Although they did not use the term *argument* per se, they emphasized the need to focus on the underlying theory regarding the phenomenon or construct of interest. When selecting measures or planning how measures will be used, they advised researchers to consider nine characteristics of the underlying theory and constructs: concreteness, type of measure, perspective, realism, level, biases, sensitivity, ideographic versus nomothetic, and dimensionality. In a sense, then, the researchers would build an argument for the definitions of the constructs.

Third, emphasizing the fact that validation is an extended, multifaceted undertaking—and not just a collection of tools or an array of correlations—helps students understand the complex nature of validation. According to Kane (2001),

> [The] validation of an interpretation in terms of a theoretical construct [involves] an extended effort, including the development of a theory, the development of measurement procedures thought to reflect (directly or indirectly) some of the constructs in the theory, the development of specific hypotheses based on the theory, and the testing of these hypotheses against observations. (p. 323)

Asking students to design validity studies for real or hypothetical measures (or ones that they design themselves, if such an activity is part of the course) can help them understand the need for many types of validity evidence. Published journal articles that describe the development and psychometric study of various kinds of instruments are quite easy to locate; indeed, Buboltz, Miller, and Williams (1999) conducted a content analysis of articles published in the *Journal of Counseling Psychology* between 1973 and 1998 and reported that articles on the development and evaluation of tests and measures ranked 4th (out of 14 categories of types of articles) in frequency of appearance. Interesting constructs can be located in research and measurement journals in counseling and psychology, and these constructs can then become the basis for class activities on developing definitions of constructs, constructing instruments, and designing validity studies. A perusal of recent editions of a few such journals produced the following examples of constructs that became the foci for measure development and validation research: mathematics anxiety (Beasley, Long, & Natali, 2001), Black identity (Cokley & Helm, 2001), self-efficacy (Fouad, Smith, & Enochs, 1997), motivational orientations toward work (Loo, 2001), expectations about counseling (Ægisdóttir, Gerstein, & Gridley, 2000), and "color-blind" racial attitudes (Neville, Lilly, Duran, Lee, & Browne, 2000).

Another type of measure that counseling students can certainly relate to when discussing the need for various types of validity evidence is a certification or licensing examination (such as the NCE). Drawing upon some suggestions offered by Hattie, Jaeger, and Bond (1999), instructors can discuss the need for evidence that the content of the test is appropriate for its use, evidence that it predicts some valued outcome (such as job performance), evidence that scores are consistent with underlying assumptions about the abilities that the test claims to measure, evidence that the scores are unbiased, evidence that the internal structure of the test matches a blueprint for the test, and so forth.

A final suggestion is to include general guidelines about ways to discuss and write about validation and validity evidence (Goodwin, 2002b). These include referring to validity as estimated, not proven or established; not describing it as a property of a test or measure per se; including details about empirical studies, such as descriptions of the subjects, testing conditions, results; and downplaying the statistical significance of correlations as compared with the practical meaning of the coefficients.

SUMMARY AND CONCLUSION

Evolutionary changes in the meaning of validity have occurred since the 1940s. The newest edition of the *Standards* (AERA, APA, & NCME, 1999) contains a dramatic shift in the definition and description of validity: the elimination of the content, criterion-related, and construct types of validity. This older tripartite or trinity view of validity has been replaced with one that focuses on five distinct types of validity evidence: evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on the consequences of testing. This is a very welcome change. The tripartite view of validity has been problematic for a number of reasons. It tends to mask the unitary nature of validity, it compartmentalizes our thinking about validity, it is incomplete (i.e., it ignores important questions about the consequences of testing), and it promotes the incorrect notion that all of the types are equal. Approaching validity as multidimensional and complex—requiring a wide and diverse body of evidence—is much more realistic and appropriate. The new way of defining validity will likely affect the field of measurement in a number of ways, especially in terms of how validity is conceptualized and estimated by test developers and researchers.

Most measurement textbooks, test manuals, and licensing/certification examinations continue to present the tripartite view of validity. Changing the treatment of validity in these and other materials (e.g., research methods textbooks) likely will be a slow process. Instructors of measurement courses are faced with the challenge of incorporating both the old tripartite view of validity and the newer, evidence-oriented meaning of validity into their courses. Suggestions for instructors of these courses include adopting the most recent *Standards* as a required or optional book for the course; emphasizing the role of arguments in validation; and helping students design validity studies for real or hypothetical measures so they gain an appreciation for the complex, extended, multidimensional nature of validation research.

REFERENCES

Ægisdóttir, S., Gerstein, L. H., & Gridley, B. E. (2000). The factorial structure of the Expectations About Counseling Questionnaire–Brief Form: Some serious questions. *Measurement and Evaluation in Counseling and Development, 33,* 3–20.

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

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

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Author.

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

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19–32). Hillsdale, NJ: Erlbaum.

Beasley, T. M., Long, J. D., & Natali, M. (2001). A confirmatory factor analysis of the Mathematics Anxiety Scale for Children. *Measurement and Evaluation in Counseling and Development, 34,* 14–26.

Berk, R. A. (1990). Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research, 12,* 659–671.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*(1), 5–9, 30.

Buboltz, W. C., Miller, M., & Williams, D. J. (1999). Content analysis of research in the *Journal of Counseling Psychology. Journal of Counseling Psychology, 46,* 496–503.

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

Chudowsky, N., & Behuniak, P. (1998). Using focus groups to examine the consequential aspect of validity. *Educational Measurement: Issues and Practice, 17*(4), 28–38.
Cokley, K. O., & Helm, K. (2001). Testing the construct validity of scores on the Multidimensional Inventory of Black Identity. *Measurement and Evaluation in Counseling and Development, 34,* 80–95.
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 & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.
Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.
Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.
Dwyer, C. A. (2000). Excerpt from validity: Theory into practice. *The Score, 22*(4), 6–7.
Fouad, N. A., Smith, P. L., & Enochs, L. (1997). Reliability and validity evidence for the Middle School Self-Efficacy Scale. *Measurement and Evaluation in Counseling and Development, 30,* 17–31.
Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist, 27,* 197–222.
Goodwin, L. D. (1997). Changing conceptions of measurement validity. *Journal of Nursing Education, 36,* 102–107.
Goodwin, L. D. (1999). The role of factor analysis in the estimation of construct validity. *Measurement in Physical Education and Exercise Science, 3,* 85–100.
Goodwin, L. D. (2002a). Changing conceptions of measurement validity: An update on the new *standards. Journal of Nursing Education, 41,* 100–106.
Goodwin, L. D. (2002b). The meaning of validity. *Journal of Pediatric Gastroenterology and Nutrition, 35,* 6–7.
Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health, 20,* 269–274.
Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6,* 427–438.
Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.
Haertel, E. H. (1999). Validity arguments for high-stake testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.
Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 393–446). Washington, DC: American Educational Research Association.
Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.
Kane, M. T. (1994). Validating interpretative arguments for licensure and certification examinations. *Evaluation & the Health Professions, 17,* 133–159.
Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342.
Langenfeld, T. E., & Crocker, L. M. (1994). The evolution of validity theory: Public school testing, the courts, and incompatible interpretations. *Educational Assessment, 2,* 149–165.
Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14.
Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14–16.
Loo, R. (2001). Motivational orientations toward work: An evaluation of the Work Preference Inventory (Student Form). *Measurement and Evaluation in Counseling and Development, 33,* 222–233.
Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16–18.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.
Messick, S. (1988). The once and future uses of validity: Assessing the meaning and consequences of validity. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.
Messick, S. (1989a). Meaning and value in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.
Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.
Millman, J. (1998). Strategies for identifying a research topic in educational measurement. *Educational Measurement: Issues and Practice, 17*(2), 37–39.
Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229–258.
Neville, H. A., Lilly, R. L., Duran, G., Lee, R. M., & Browne, L. (2000). Construction and initial validation of the Color-Blind Racial Attitudes Scale (CoBRAS). *Journal of Counseling Psychology, 47*, 59–70.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.
Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.
Shepard, L. A. (1993). Evaluating test validity. In Darling-Hammond (Ed.), *Review of research in education* (vol. 19, pp. 405–450). Washington, DC: American Educational Research Association.
Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8, 13, 24.
Silverlake, A. C. (1999). *Comprehending test manuals: A guide and workbook*. Los Angeles: Pyrczak Publishing.
Sireci, S. G., & Green, P. C. (2000). Legal and psychometric criteria for evaluating teacher certification tests. *Educational Measurement: Issues and Practice, 19*(1), 22–31, 34.
Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Merrill.
Tracey, T. J. G., & Glidden-Tracey, C. E. (1999). Integration of theory, research design, measurement, and analysis: Toward a reasoned argument. *The Counseling Psychologist, 27*, 299–324.

APPENDIX

List of Textbooks Reviewed

Aiken, L. R. (2003). *Psychological testing and assessment* (11th ed.). Boston: Allyn & Bacon.
Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart, & Winston.
Drummond, R. J. (1996). *Appraisal procedures for counselors and helping professionals* (3rd ed.). Englewood Cliffs, NJ: Merrill.
Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Needham Heights, MA: Allyn & Bacon.
Kubiszyn, T., & Borich, G. (2000). *Educational testing and measurement: Classroom application and practice* (6th ed.). New York: Wiley.
Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Merrill.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.
Whiston, S. C. (2000). *Principles and applications of assessment in counseling*. Australia: Brooks/Cole.
Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and assessment in schools* (2nd ed.). New York: Longman.