Changing Conceptions of Measurement Validity: An Update on the New *Standards*

Laura D. Goodwin, PhD

ABSTRACT

This article serves as a follow up to a 1997 article in the *Journal of Nursing Education,* in which the author presented a historical overview of the ways in which views of measurement validity had changed during the past half century. The new American Educational Research Association, American Psychological Association, and National Council on Measurement in Education *Standards for Educational and Psychological Testing,* released in 1999, includes a revised conceptualization of validity. The key changes, including the elimination of the old "trinity" view of validity and the operationalization of validity as five types of evidence, are described in this article, and specific ways to obtain evidence of each type are provided. The article concludes with a brief discussion of some of the major continuing issues and challenges in validity theory and practice.

A few years ago, the author presented a historical overview of the ways in which measurement validity definitions and estimation techniques had changed during the past half century (Goodwin, 1997). 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), which is only the sixth revision of this document since the original *Technical Recommendations for Psychological Tests and Diagnostic Techniques* was published in 1954 (APA, 1954), was released in 1999. The purpose of this article is to provide an update on key elements in the new *Standards* that pertain to the meaning of measurement validity and recommendations for its estimation. As will be shown, there are several important changes in the way validity is conceptualized and described.

In her 2000 presidential address to the APA's Division 5 (evaluation, measurement, and statistics) members, Dwyer (2000) described four obstacles to moving validation theory into practice. One of the obstacles she described was "knowing what validity is today":

> The key writings on validation are not accessible to busy practitioners. They are written in ways that facilitate the development of theory, but do not speak directly to the concerns of practitioners as they experience them. In other words, they are hard to understand, and it is not clear exactly how they pertain to practitioners' problems.
>
> There are important documents that should overcome this obstacle—but they do so only partially. The *Standards for Educational and Psychological Testing* are not—nor are they intended to be—a cookbook. They leave a great deal of room for interpretation… (p. 6).

The aim of this article is to help practitioners (educators and researchers) interpret the new *Standards* by operationalizing the types of validity evidence presented therein. The main differences in the definitions or the treatment of the types of evidence between the 1999 and 1985 editions of the *Standards* also will be highlighted. The article will conclude with a description of some of the continuing issues and challenges for validity theory and practice.

BACKGROUND

As many authors on the history of measurement validity have noted (e.g., Geisinger, 1992; Goodwin, 1997,

Dr. Goodwin is Professor, University of Colorado at Denver, School of Education, Denver, Colorado.

Address correspondence to Laura D. Goodwin, PhD, Professor, University of Colorado at Denver, School of Education, CB106, PO Box 173364, Denver, CO 80217; e-mail: Laura_Goodwin@ceo.cudenver.edu.

1999; Langenfeld & Crocker, 1994; Messick, 1989b; Moss, 1992; Shepard, 1993), one of the major developments in the evolution of validation theory and practice was the categorization of validity into three types by Cronbach and Meehl (1955):

- Content validity.
- Criterion-related validity, which included two subtypes—concurrent validity and predictive validity.
- Construct validity.

This holy trinity (Guion, 1980) view of validity appeared in the following edition of the *Standards for Educational and Psychological Tests and Manuals* (APA, AERA, & NCME, 1966) and subsequently in many (if not virtually all) measurement textbooks. Although this way of conceptualizing validity was useful to measurement theorists and practitioners for many years, it also caused some problems and confusion. It tended to compartmentalize thinking about validity, narrowing or limiting it to a checklist type of approach. It made it easy to overlook (consciously or unconsciously) the fact that construct validity is really the "whole" of validity theory (Shepard, 1993); it interfered with acceptance of the very reasonable notion that validity is a unitary concept (Geisinger, 1992); and it encouraged the misconception that validity is somehow a property of a test or measure per se, rather than a property of the scores obtained with a measure when used for a specific purpose and with a particular group of respondents (Goodwin & Goodwin, 1999).

Having taught a graduate-level measurement course for the past 20 years to students in nursing and education programs, the author knows firsthand how confusing the "breakdown" approach to validity can be. Instead of spending precious time on the larger and more important issues about validity, educators often become bogged down in the classification of validity studies (real and hypothetical) into the three or four specific types. Although the 1985 *Standards* (AERA, APA, & NCME, 1985) cautioned that "the use of the category labels should not be taken to imply that there are distinct types of validity" (p. 9), this conceptualization of validity had been used and taught for so long that it was difficult not to rely on it. One of the major changes in the new *Standards* (AERA, APA, & NCME, 1999)—the omission of the "trinity" view of validity—is a welcome change.

The criticality of choosing, developing, and using tests and other measures wisely, based on adequate evidence of validity of scores obtained under certain conditions and for particular purposes, cannot be understated. As written in the introductory section of the 1999 *Standards:*

> The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use and also can provide a route to broader and more equitable access to education and employment. The improper use of tests, however, can cause considerable harm to test takers and other parties affected by test-based decisions. The intent of the *Standards* is to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices (AERA, APA, & NCME, 1999, p. 1).

Before describing the major changes in the 1999 *Standards'* treatment of types of validity evidence, it is relevant to briefly and generally point out the major content differences between the 1999 edition and the 1985 *Standards* (AERA, APA, & NCME, 1985). First, the total number of standards has increased from 180 to 264. Reasons given for the additional standards include new developments in measurement (e.g., the increased use of performance assessments, greater focus on the consequences of test use), as well as the need to add standards on nontechnical issues such as avoiding conflicts of interest and treating all test takers fairly. Second, the designation of each standard as either primary, secondary, or conditional, to reflect varying levels of importance, has been eliminated. Instead, the text that accompanies each standard now includes a discussion of the conditions under which the standard is relevant. Third, the definition of the term construct has been altered. Historically, this term referred to an unobservable characteristic, one which is inferred from observations or scores obtained with a test or other measure. Currently, construct is defined in a broader way, as a characteristic or concept that a test or other measurement procedure is intended to measure. Finally, each chapter includes more introductory text, providing background information for the standards contained therein, to enhance the usefulness of the *Standards* in educational and training programs for future test developers and users.

TYPES OF VALIDITY EVIDENCE AS DESCRIBED IN THE NEW *STANDARDS*

In the chapter on validity in the 1999 *Standards,* validity is defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p. 9). It is described as the most fundamental concern in the development and evaluation of measures, with the process of validation being the accumulation of evidence to provide a sound scientific basis for proposed interpretations of scores. Further, an important and often misunderstood fact is stated clearly, "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 much be validated" (AERA, APA, & NCME, 1999, p. 9).

Prior to the presentation of the 24 validity standards, descriptions of five distinct types of validity evidence are provided:

- Evidence based on test content.
- Evidence based on response processes.
- Evidence based on internal structure.
- Evidence based on relations to other variables.
- Evidence based on the consequences of testing.

Two important points are made preceding and immediately following the descriptions of these types of validity

evidence. One highlights the omission of the former breakdown of validity into three or four types:

> 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 purpose (AERA, APA, & NCME, 1999, p. 11).

The second important point emphasizes the need to accumulate and integrate validity evidence:

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. It encompasses evidence gathered from new studies and evidence available from earlier reported research. The validity argument may indicate the need for refining the definition of the construct, may suggest revisions in the test or other aspects of the testing process, and may indicate areas needing further study (AERA, APA, & NCME, 1999, p. 17).

The continuing evolution of the meaning of measurement validity is demonstrated well by an examination of the five types of validity evidence described in the 1999 *Standards.* In the Table, each of the five types of evidence is listed, along with examples of specific validation activities that can be used to obtain the evidence. For comparative purposes, the Table also includes the corresponding name of the type of validity evidence as it was presented in the 1985 *Standards* when possible.

Evidence Based on Test Content

The first type of validity evidence is based on test content. This was referred to in the 1985 *Standards* as content-related evidence and commonly has been termed content validity in measurement and research methods textbooks for many years. In general, this type of evidence attempts to answer the question, "To what extent does the content of the test (including specific items or tasks and their formats) represent the content domain?" It is relevant for virtually all types of measures in both the cognitive and affective domains and for measures using various formats, including paper-and-pencil self-report measures, projective devices, observation schedules, and interview protocols. Of all of the types of validity evidence, it is the one type that pertains to the content of the measure per se, as well as to the relevance of the content domain to the proposed interpretations of scores obtained with the measure.

Logical analyses and experts' evaluations of the components of the measure are the key ways this type of validity evidence is obtained. In addition to expert review of such aspects as sufficiency, relevancy, and clarity of the components, evidence about the extent to which a test measures more or less than the intended construct is also often very important. These concerns are called construct underrepresentation and construct-irrelevant variance. Discussion of these problems sometimes occurs along with more general concerns about bias (e.g., what it is, how to detect it, how to eliminate its influence).

The importance of content validity evidence has been noted by many authors, including Sireci and Green (2000), who described its influence on court decisions emanating from challenges to the validity of certification examinations. However, it is curious that content validity evidence tends to be relegated to a lower level of importance in measurement textbooks (Berk, 1990) and that there is a general paucity of information related to how to obtain useful evidence (Slocumb & Cole, 1991). Helpful guides in this area have been offered by Berk (1990), Grant and Davis (1997), and Davis (1992). Interested readers will find information in these references on such aspects of content validation as the selection, preparation, and use of experts; the optimal number of experts; and the collection and quantification of content validity evidence data.

Evidence Based on Response Processes

The second type of validity evidence is based on response processes. Formerly, this type of evidence was one component of construct-related evidence. This type of evidence addresses the question, "To what extent does the type of performances or responses in which examinees engage fit the intended construct?" For example, if the developers of a measure claim that the measure assesses self-efficacy, evidence that respondents are not merely providing socially desirable answers would be relevant. The kinds of validation activities that can provide this type of evidence include actually analyzing individual responses via interviews with respondents, or observing respondents as they engage in task performances. Another example of a validation activity that belongs in this category is studying the ways in which judges or observers use criteria to record and score performances, essays, and behaviors. 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.

Evidence Based on Internal Structure

A third type of validity evidence is based on internal structure. This type of evidence also was included as one type of construct-related evidence in the 1985 *Standards,* and it generally answers the question, "To what extent do the relationships among test items and components match the construct as operationally defined?" Most closely associated with this kind of evidence is factor analysis, especially confirmatory factor analysis. Because evidence based on factor analysis is relatively easy to obtain, results of factor analytical studies are reported quite commonly. Although factor analysis can produce useful information, it is also only one type of information. An overreliance on factor analysis as a validation technique can lead to a narrow body of empirical support for validity arguments (Goodwin, 1999; Nunnally & Bernstein, 1994; Thorndike, 1997). A cautionary statement by Thorndike (1997) is par-

TABLE
Types of Validity Evidence as Presented in the 1999 *Standards*

Type of Validity Evidence*	Examples of Validation Activities
A. Evidence based on test content (Content-related evidence)	1. Logical analyses of extent to which the test content represents the content domain.
	2. Experts' evaluations of the extent to which the items or subparts of a test match the definition of the construct and/or the purpose of the test.
	3. Experts' evaluations of the relevance, importance, and clarity of items or tasks.
	4. Experts' evaluations of the extent to which construct underrepresentation or construct-irrelevant components may give unfair advantages to one or more subgroups of examinees.
B. Evidence based on response processes (Construct-related evidence)	1. Analyses of individual responses, via interviews with test takers.
	2. Monitoring changes in, or development of, responses to performance tasks.
	3. Process studies, examining similarities and differences in responses given by members of distinct subgroups of examinees.
	4. Studies on the ways in which judges, observers, and interviewers collect, record, and interpret data.
C. Evidence based on internal structure (Construct-related evidence)	1. Factor analytical studies.
	2. Item analyses to examine item interrelationships.
	3. Differential item functioning (DIF) studies.
D. Evidence based on relations to other variables	1. Correlational studies of the type and extent of relationships between scores and external variables.
(D1-2: Criterion-related evidence; also, concurrent and predictive validity)	2. Correlational studies of the extent to which scores forecast or predict criterion performance or scores on measures obtained at a later date.
(D3-6: Construct-related evidence)	3. Convergent validity studies, investigating the relationships between scores and other tests intended to measure similar constructs.
	4. Discriminant validity studies, investigating the relationships between scores and other measures of purportedly different constructs.
	5. Experimental studies, intended to test hypotheses about effects of specific interventions on test scores.
	6. Known-group comparison studies, intended to test hypotheses about expected differences in test scores across specific groups of examinees.
(D7-9: Criterion-related evidence)	7. Studies of the effectiveness of selection, placement, and classification decisions.
	8. Differential group prediction or relationship studies.
	9. Studies of validity generalization.
E. Evidence based on consequences of testing	1. Descriptive studies of the extent to which anticipated benefits of testing are realized.
	2. Descriptive studies of the extent to which unanticipated negative consequences occur.

* Where possible, the corresponding type of validity, as presented and discussed in the 1985 *Standards,* is given in parentheses.
Source: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999).

ticularly apt: "this internal or factorial validity still needs evidence of a relationship to life events outside the tests themselves if the factors are to have substance, vitality, and scientific or educational utility" (p. 159).

Another validation activity that fits within the internal structure category of evidence is the investigation of differential item function (DIF). Differential item function studies are conducted to detect item bias. According to Hattie, Jaeger, and Bond (1999), "...DIF is present when examinees of the same ability but belonging to different groups have differing probabilities of success on an item" (p. 432). Therefore, DIF occurs when an item parameter (e.g., discrimination index) differs across groups, such as various ethnic groups or examinees grouped by gender. Nunnally and Bernstein (1994) described and illustrated a number of approaches to the assessment of DIF, including approaches based on both item response theory and classical test theory.

One specific technique used to examine DIF is the Mantel-Haenszel procedure (Holland & Thayer, 1988;

Silverlake, 1999). In this procedure, the groups (e.g., men versus women) first are matched to form two groups that are similar in terms of their total scores on a test. Note that the groups would not include all women and men who were tested, rather only those women who can be matched with men and men who can be matched with women. The performance of each group on each item then is compared, and items that show statistically significant differences in difficulty indexes are flagged for further review for bias.

Evidence Based on Relations to Other Variables

The next type of validity evidence, that which is based on relations to other variables, includes many approaches to validation. In general, this type of evidence provides answers to such questions as:

- "What is the nature and extent of the relationships between test scores and other variables, including other variables that the test is expected to correlate with or predict, as well as other variables that the test is not expected to relate with?"
- "To what extent are these relationships consistent with the construct that underlies the proposed interpretations of scores?"

The specific approaches used to obtain this type of evidence are most commonly correlational and group-comparison studies. Formerly (i.e., the 1985 *Standards* and earlier editions, the results of these studies had several different labels and, in many ways, represented one of the greatest sources of confusion for students struggling to categorize validity studies and results according to the "trinity" view of validity.

Concurrent validity, the extent to which scores obtained with a measure correlate with scores from another measure of the same construct or trait, is a common type of validity evidence. Predictive validity, the extent to which scores obtained with a measure correlate with data obtained at a later date, is also quite typical. Together, these two types of validity evidence have been called criterion-related validity for many years. Also part of this category of validity evidence, and presented as part of criterion-related validity in the 1985 *Standards,* are the results of studies of the accuracy of selection, placement, and classification decisions made on the basis of test scores; differential group prediction studies; and studies of validity generalization. The latter represents an important issue in employment testing (i.e., the extent to which evidence of validity, usually predictive in nature, obtained in one setting can be generalized to other settings) (Schmidt & Hunter, 1977). Meta-analysis (Cooper & Hedges, 1994; Hedges & Olkin, 1985) is a useful technique for the investigation of validity generalization but can be used only if a sufficient database of findings from specific validity studies is available to warrant its use.

Other types of validity evidence based on examining relations with other variables are convergent and discriminant validity evidence, traditionally associated with construct-related validity. Convergent validity evidence indicates the relationships between test scores and scores from other measures aimed at assessing similar constructs. Discriminant validity evidence indicates relationships between the scores and other measures intended to assess different constructs—ones that, theoretically, should not correlate with the scores from the measure being validated. Campbell and Fiske (1959), who proposed the multitrait-multimethod (MTMM) approach to construct validation, coined these terms. Studying the relationships among different methods of measurement of the same construct, which can help sharpen the meaning of the construct, is also part of the MTMM model. Finally, the Table lists both experimental and known-group comparison studies as ways to examine relationships between test scores and other variables. These are usually hypothesis driven, with the hypotheses emanating from the theoretical and/or operational definitions of the constructs. Such studies traditionally also have been considered part of construct-related validity evidence.

Evidence Based on the Consequences of Testing

The fifth type of validity evidence described in the 1999 *Standards* is evidence based on the consequences of giving tests, administering other kinds of measures, collecting data with observations or interviews, and so forth. The questions answered with this type of evidence are:

- "To what extent are anticipated benefits of measurement realized?"
- "To what extent do unanticipated benefits (positive and negative) occur?"
- "To what extent are differential consequences observed for different identifiable subgroups of examinees?"

Studying the consequential aspects of testing hardly was mentioned in the 1985 edition and was not listed as a separate type of validity evidence.

Dwyer's (2000) fourth obstacle to moving validation theory to practice was related to consequences. She pointed out that the measurement community has had great difficulty accepting Messick's (1989a, 1989b) rationale for including consequential aspects of validity as part of the validation process because the study of these aspects goes beyond traditional psychometric boundaries into policy areas. In 1997, prominent measurement experts debated the topic in a special issue of *Educational Measurement: Issues and Practice* (Linn, 1997; Mehrens, 1997; Popham, 1997; Shepard, 1997). In the editorial that preceded the debate, Crocker (1997) summarized the key issue:

> The question is whether the investigation of possible consequences of administering and using an assessment should be defined as an integral part of the validation plan. For many, the issue is whether validation is to be regarded as a scientific, empirical enterprise or a sociopolitical process as well. Ultimately, the prevailing argument in this debate will shape the nature of measurement practice and professional preparation for years to come (p. 4).

Perhaps because of the relative recency with which the idea of consequential validity was introduced, and the

ongoing controversy about whether it even belongs in validation theory and practice, there have been few actual suggestions on how it should be estimated. Chudowsky and Behuniak (1998) offered an example of the use of focus groups to examine the consequential validity of a large-scale assessment program. Millman (1998) noted the large amount of work that needs to be completed in this area and posed some specific questions, such as, "What consequences are we talking about? If only intentional consequences then whose intentions? The users? Which users? The test sponsors? The test developers? What should a consequential validity study look like?" (p. 38).

CONTINUING ISSUES AND CHALLENGES IN VALIDATION THEORY AND PRACTICE

In addition to many outstanding questions about the fifth type of validity evidence (i.e., the consequential aspects of validity), there are a number of other important issues and challenges in validation theory and practice that are not resolved in the 1999 *Standards.* One issue relates to the nature of construct validity. Although construct validity currently is synonymous with validity itself, there is a "problem" with it, as recently noted by Fremer (2000), "We [in the *Standards*] have elevated the concept of construct validation to so high a level that it seems an 'out of reach' goal" (p. 1). Dwyer (2000) also included this theme in her presidential address, when she spoke about the disparate natures of the worlds of practice and research:

> I have observed that many practitioners fear the very idea of dealing with construct validity—feel that the ongoing, argument-based nature of it is impossible to fulfill in practical circumstances, or even that they will appear foolish in the world of theory if the evidence they offer up is not good enough. In a way, there was comfort in numbers, when a high correlation coefficient might be all that was needed to support the quality of a test. It is now widely known that this is not enough; but less widely understood is that it is not even the right way to think about the question (pp. 6-7).

Many difficult questions exist about validity:

- Exactly how does a researcher develop a convincing validity argument for a particular measure?
- How much evidence needs to be accumulated?
- To what extent can the findings from just one (or just a few) validity studies be generalized to other settings, populations, purposes for the measurement, and so forth?

Although these are not new questions, the revised views of appropriate and important types of validity evidence that appear in the 1999 *Standards* should promote some fresh thinking about the "how to's" of validation research. For example, Fremer (2000) reported that The College Board has begun a project to develop a Joint Committee on Testing Practices Casebook centered on the *Standards.*

An even more fundamental question that many likely would ask is, "Just what is a 'validity argument?" Haertel (1999), in his presidential address at the annual meeting of the National Council on Measurement in Education, described his view:

> The idea of *validity argument*…is that validation should be a process of constructing and evaluating arguments for and against proposed test interpretations and uses. 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…. Now that [assembling data in support of both sides of the argument] is a real challenge. Instead of being advocates for the work we do, we are asked to embrace the scientific ideal of striving to understand and disclose the weaknesses as well as the strengths of our tests and testing practices (pp. 5-6).

The question of bias detection constitutes another interesting aspect of the 1999 *Standards.* While the study of culture, gender, and age bias in scores long has been recognized as a critical aspect of validation research, points that bear on this topic are scattered throughout the presentation of the five types of validity evidence. In the Table, examples of validation activities that directly pertain to bias detection are those described in points A4, B3, C3, and D8. Other validation efforts, including the study of the consequences of using measures (E), certainly will often pertain to the issue of bias, as well. However, differentiating between adverse impacts and bias is an important and challenging concern (Hattie et al., 1999).

Finally, there are several continuing themes in terms of validation that should be reiterated or stressed. One is the inappropriateness of saying such things as "this test is valid" or referring to "the validity of this measure," as though validity were some sort of static property of a test or other measure. It is the inferences drawn from the data obtained by using a particular measure that have some degree of validity. Relatedly, the characteristics of the sample used in a particular validation study always should be completely described. It is important to learn as much as possible about them, including the amount of variability in their scores. The latter is especially important in correlational validity studies because the heterogeneity of the scores of specific subjects will affect the results. In addition, the measurement theory or model used to derive a validity estimate may affect the resulting coefficient, just as it does in reliability estimation (Thompson & Vacha-Haase, 2000), so it is important that the theory or model be specified and that interpreters do not assume that other models would yield the same result.

CONCLUSIONS

The new *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) are a valuable tool for nursing researchers and practitioners. One useful component is the revised conceptualization of types of validity evidence, and it is particularly helpful that the

old "trinity" view of validity has been eliminated. However, it will take time for this way of thinking about validity to become commonplace. New and revised versions of research and measurement textbooks, as well as the orientations adopted by instructors of courses that use those books and other resources, will play key roles in affecting changes in how validity is operationally defined. Hopefully, the 1999 *Standards* will stimulate new and continued efforts to study and strengthen measurement practices in nursing.

REFERENCES

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

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

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

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.

Cooper, H., & Hedges, L.V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Crocker, L. (1997). Editorial: The great validity debate. *Educational Measurement: Issues and Practice, 16*(2), 4.

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

Davis, L.L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5,* 194-197.

Dwyer, C.A. (2000). Excerpt from validity: Theory into practice. *The Score, 22*(4), 6-7.

Fremer, J. (2000). Promoting high standards and the "problem" with construct validation. *National Council on Measurement in Education Newsletter, 8*(3), 1.

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., & Goodwin, W.L. (1999). Measurement myths and misconceptions. *School Psychology Quarterly, 14,* 408-427.

Grant, J.S., & Davis, L.L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health, 20,* 269-274.

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

Haertel, E.H. (1999). Validity arguments for high-stakes 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.

Hedges, L.B., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

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.

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

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

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 and Macmillan.

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.

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. *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.

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.

Slocumb, E.M., & Cole, F.L. (1991). A practical approach to content validation. *Applied Nursing Research, 4,* 192-195.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174-195.

Thorndike, R.M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Merrill.