HKJOT 2010;20(1):30–42

REVIEW ARTICLE

Construct Validity: A Unitary Concept for Occupational Therapy Assessment and Measurement

Ted Brown

Background: In the traditional view of validity, it is divided it into three sub-types: content, criterion-related, and construct. Validity has recently been reconceptualized as a unitary factor known as construct validity. Five sources of construct validity evidence have been specified: test content, response processes, internal structure, relationships to other variables, and consequential aspects of construct validity. They function either as general validity criteria or as standards for all assessment and measurement.
Purpose: The purpose of the manuscript is to present an overview of the contemporary conceptualization of construct validity and its relevance to occupational therapy practice, education, and research.
Key Issues & Implications: Taken together, the five components of construct validity evidence provide a way of addressing the multiple and interrelated validity questions that need to be answered in order to justify occupational therapy test score interpretation and use by practitioners. Given the context of evidence-based practice, professional standards of practice, increasing calls for accountability, and the fact that validity of a test, instrument, or scale is now seen as being a dynamic process, it is important for occupational therapists to be conversant with this contemporary view of construct validity, and the body of validity evidence related to the assessment tools used in clinical practice.

KEY WORDS: Assessment • Construct validity • Measurement • Testing • Occupational therapy

Introduction

Occupational therapists often use standardized tests, instruments, and scales as part of their assessment, goal setting, treatment planning, intervention process, and follow-up evaluation when working with clients (Asher, 2007; Benson & Schnell, 1997; Craik, Davis, & Polatajko, 2007; Cronbach, 1988; Fawcett, 2007; Hinojosa & Kramer, 1998; Law, Baum, & Dunn, 2005). To ensure that clinical tests are accurately assessing what they purport to measure, they must demonstrate evidence of reliability and validity (Anastasi & Urbina, 1997; Kielhofner, 2006; Streiner & Norman, 1995). Traditionally, validity was viewed as a three-part concept that comprised content, criterion-related, and construct validity (Anastasi, 1986, 1988; Angoff, 1988; Geisinger, 1992; Law & Baum, 2005; Nunnally & Bernstein, 1994; Portney & Watkins, 2000; Yun & Ulrich, 2002). Recently, validity has been reconceptualized as being a unitary concept now known as construct validity (Downing, 2004; Kane, Crooks, & Cohen, 1999; Markus, 1998; Messick, 1989). In this contemporary context, validity refers to evidence generated to support or refute the meaning or interpretation assigned to results generated by a test, instrument or scale (Goodwin &

Department of Occupational Therapy, School of Primary Health Care, Faculty of Medicine, Nursing and Health Sciences, Monash University—Peninsula Campus, Frankston, Victoria, Australia.
Reprint requests and correspondence to: Dr. Ted Brown, Department of Occupational Therapy, School of Primary Health Care, Faculty of Medicine, Nursing and Health Sciences, Monash University—Peninsula Campus, Building G, 4th floor, McMahons Road, Frankston, Victoria, 3199, Australia.
E-mail: ted.brown@med.monash.edu.au

Leech, 2003). "Validity is never assumed and is an ongoing process of hypothesis generation, data collection and testing, critical evaluation and logical inference" (Downing, 2003, p. 831).

It is essential that occupational therapists are cognizant of and conversant with this contemporary view of construct validity since it directly impacts their daily practice with the clients they serve. The purpose of this paper is to describe the new meaning and framing of construct validity and its implications for occupational therapy practice, assessment, research, and measurement. Initially, a brief historical review of the concept of validity will be provided and then the new unitary view of construct validity will be discussed in occupational therapy contexts. Key points will be illustrated with practical examples for the reader.

Literature Review

Evolution of Validity as a Concept

The concept of validity has evolved over the last half century (Jonson & Plake, 1998; Shepard, 1993). Historically, the methods of establishing a test's validity were founded on the early work of Cronbach (1971, 1988, 1989), Cronbach and Meehl (1955), and Kane (1992, 1994, 2001). In the 1940s, validity focused on the test itself and was conceptualized as a static property of an instrument. In other words, once the validity of an instrument or scale was established, it did not change nor, was it affected by the context where the test was given or the traits of the test-takers.

The 1954 edition of the American Psychological Association's (APA) *Technical Recommendations for Psychological Tests and Diagnostic Techniques* outlined four types of validity, each relating to different inferences depending on the purpose of testing (APA, 1954). The first, content validity, dealt with the selection of items from a universe of items, and the evaluation of the extent to which the test items represented the theoretical construct of interest. For example, if a test assessed a child's visual perceptual skills, then the items on that test needed to be representative enough of the content domain of children's visual perceptual abilities.

Predictive validity was the second type. It dealt with the ability of an instrument to predict the future performance of respondents on some specific skill or area of knowledge. For example, whether the performance of children attending kindergarten on a test of visual-motor integration skills could accurately predict whether or not the children would have difficulty with cursive writing skills in Grade 3. Concurrent validity, the third type, was considered when a new instrument was proposed as a substitute or replacement for some less convenient measure that had already been accepted as the standard for use in the field. For example, whether a newly developed test of sensory integration and motor skills could replace the traditionally used Sensory Integration and Praxis Test. The final type, construct validity, was considered essential when inferences were to be made about latent (unobservable) traits. In other words, the items of a test should adequately represent the theoretical construct being assessed by a test, instrument, or scale. For example, the items on the *Bruininks-Oseretsky Test of Motor Proficiency, 2nd edition* (BOT-2) need to be representative of the fine and gross motor skill constructs it purports to assess.

At this time, Campbell and Fiske (1959) promoted the idea of discrete forms of validity and the need for multiple kinds of validity evidence in their seminal paper introducing the multimethod-multitrait approach to validation. This included the introduction of convergent, divergent, diagnostic, and discriminant types of validity under the construct validity category. In the 1960s, the meaning of validity shifted in focus to use. In other words, validity was defined as the extent to which a test generated information that was helpful for a specific purpose or use.

The 1966 version of the *Standards for Educational and Psychological Tests and Manuals* (*Standards*) combined concurrent validity and predictive validity into criterion-related validity (APA, American Educational Research Association [AERA], & National Council on Measurement in Education [NCME], 1966). The 1966 edition of the *Standards* included the traditional "holy trinity" view of validity by categorizing it into specific types: content validity, criterion-related validity (which included concurrent and predictive validities), and construct validity (see Figure 1; Cronbach & Meehl, 1955; Guion, 1980). Both the 1954 and 1966 editions of the *Standards* linked the four types of validity to particular aims of testing.

Even though this way of conceptualizing validity was helpful to measurement theorists, educators, researchers and practitioners, it also led to notable challenges and confusion. This three-type validity view tended to compartmentalize the thinking about validity, thus narrowing or limiting it to a checklist approach (e.g. a test user merely checking off the types of validity that were reported in a test's manual) (Goodwin, 1997, 2002a), and is what many occupational therapists still refer to when evaluating research studies in the context of

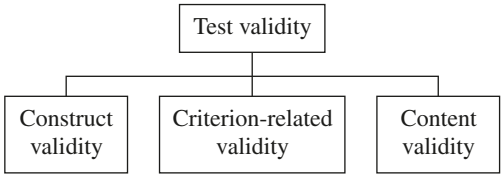

Figure 1. *Traditional tripartite view of test validity.*

evidence-based practice or when examining a new assessment tool for potential use with clients in clinical contexts (Benson & Schnell, 1997; Kielhofner, 2006). For example, when considering the effectiveness of clinical interventions based on sensory integration theory, many occupational therapists would report information about the content, criterion-related, and construct validity of the assessment tools used to gather the data as one means of critiquing the quality and/or level of evidence of the research reported. However, the construct validity of an assessment tool is never static, but always dynamic and cumulative.

The 1985 *Standards for Educational and Psychological Testing* (*Standards*) was essentially unchanged from its earlier versions, except that it provided more detail about what type of evidence was appropriate for each type of validity (AERA, APA, & NCME, 1985). However, another shift in the conceptualization of validity was taking place with psychometricians and testing experts emphasizing the inferences and decisions made from test scores (Cronbach, 1988, 1989; Messick 1988, 1989). For example, test scores are frequently used as one selection criteria for students at the tertiary and postgraduate level. Students applying for admission to medical school often complete the standardized *Medical College Admission Test* (MCAT) while students applying for masters and doctoral programs often complete the *Graduate Record Examination* (GRE). The inference that is made from the MCAT and GRE scores is that students with higher performance scores are more able to succeed in their chosen fields of study. The MCAT and GRE scores students receive also impact the decisions made by universities and colleges about offers of student admission and scholarship eligibility. The inferences and decisions made from the MCAT and GRE scores, for example, are that they are used in a competitive way to allocate limited educational resources.

Two further shifts in the meaning of validity were also taking place. First, the usefulness and relevance of the traditional trinity view of validity was being challenged and second, validity was conceptualized as a unitary concept with "construct validity" being the key and unifying type of validity (Kane, 1994; Langenfeld & Crocker, 1994). One of the primary reasons for the shift in the view of validity from a "tripartite" to a "unitary" concept was that validity theorists promoted the idea that the three tier approach (content, criterion-related, and construct) was artificial and that a body of dynamic validity evidence was required for tests, instruments and scales.

One final issue that was gaining attention in the validity arena was the need for evidence about the social consequences of test use. "What was new (and controversial) in discussions about the role of consequences in validation research was studying both the intended and unintended—often adverse—consequences of test use" (Goodwin & Leech, 2003, p. 182). The issue of the "social consequences of test use" has relevance for occupational therapists. For example, test scores may be used to determine whether a child is eligible for certain types of private or public funding for therapy services or specialized equipment. What are the consequences for a child (and his/her family) who obtains a test score above the funding threshold cutoff and is discharged prematurely, when in fact it was the test that was problematic and did not fairly or realistically evaluate a child's performance? When a test is not sensitive, not accurate, or biased, it can cause potentially negative social, psychological, and economic consequences for the test-takers who are evaluated. For example, if students from culturally diverse backgrounds or lower socioeconomic backgrounds write an achievement test to be eligible for university scholarships, but the achievement test was designed for middle-class, Caucasian students, then the students writing the exam would be disadvantaged since the test is biased.

In the most recent edition of the *Standards* (AERA, APA, & NCME, 1999), the conceptualization of validity markedly changed from what it was in its three earlier editions (AERA, APA, & NCME, 1985; APA, 1954; APA, AERA, & NCME, 1966). The view of validity theory prevailing today is largely based on the seminal work of Messick (1989, 1995). The current emphasis states that all validity is subsumed under construct validity and is concerned with "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other models of assessment" (Messick, 1989, p. 741).

Although there are numerous and multidimensional methods available to determine validity, validity is now viewed as a unitary concept. The various approaches to it are related components that can be combined to evaluate what inferences can be made from test scores and test-taker performance results (Jonson & Plake, 1998; Smith, 2001). In the 1999 *Standards*, validity was 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). The most important issue in the development and evaluation of tests, instruments, scales, and measures is the process of validation, which involves the accumulation of evidence to provide a sound empirical foundation for the proposed interpretations of test scores. Downing (2003) eloquently states this as "validity requires an evidentiary chain which clearly links the interpretation of the assessment scores or data to a network of theory, hypotheses and logic which are presented to support or refute the reasonableness of the desired interpretations" (p. 831).

Although the unitary concept of construct validity published in the 1999 *Standards* has been widely endorsed (Goodwin & Leech, 2003; Kane, 2001; Wolfe & Smith, 2007), two validity theorists, Lissitz and Samuelsen (2007), have recently proposed further revisions to the concept of validity. They have articulated a number of criticisms on the "unifying" notion of construct validity and proposed a different method of conceptualizing how validity can be established. Lissitz and Samuelsen propose that validity should be established by considering whether the focus of the investigation of a test is *internal* to the test itself or focuses on constructs and relationships *external* to the test.

In other words, test evaluation is separated into internal and external aspects; the test validator first looks at the internal aspects of a test and then moves to the external aspects. In their view, the *internal validity* of a test is established by investigating the test's reliability and content validity and that other characteristics, such as criterion-related and construct validity are viewed as aspects of *external validity* (Lissitz & Samuelsen, 2007). Lissitz and Samuelsen's proposal for changes to the concept of validity, has not as yet been widely accepted and has also been criticized by a number of other theorists such as Gorin (2007), Mislevy (2007) and Sireci (2007).

Problems With the Tripartite View of Validity

The traditional concept of validity, as previously mentioned, divides it into three separate subtypes: content, criterion-related, and construct validity (AERA, APA, & NCME, 1985; see Figure 1). This tripartite view of validity is problematic for a number of reasons. The categorization of validity into three distinct types of validity is, for example, frequently confusing for students, since the difference between construct validity and the other types of validity is not overt. For example, the difference between concurrent validity (a subtype of criterion-related validity) and convergent validity (a subtype of concurrent validity) is not always clear.

A second issue raised by Goodwin and Leech (2003) is that the former three-level view of validity encouraged a "hierarchical checklist approach" to validity when students were critiquing the existing validity evidence of specific tests, instruments, and scales. A third issue is that the tripartite validity categories promote a misconception that validity is a static property of a test, instead of being influenced by the respondent sample. For example, the validity results of a test, scale, or instrument are often influenced by the context of the environment where the test is completed, the traits of the test-takers (e.g. age, gender, socioeconomic level, geographical region of residence, family constellation), and the method of participant recruitment and sampling used (e.g. convenience sampling, paid participants, random selection of participants). A final issue is that the tripartite view of validity runs counter to the ideas of the "whole" of validity theory (Downing, 2003).

Messick (1989) considered the traditional tripartite view of validity to be fragmented and incomplete since it failed to take into account both the evidence of the value implications of score meaning as a basis for action and the social consequences of score use (e.g. did respondents benefit or experience negative consequences for completing a test). In his view then, validity was not so much a property of a scale as it was the meaning of the scale scores. In the traditional conception, test scores were not only based on the scale items, but also on the participants' responses to the test items and the context of the assessment. But, according to Messick (1989), what needed to be valid was the meaning or interpretation of test scores, as well as the implications for actions that this meaning encompassed. "The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question" (Messick, 1994, p. 741). This is the primary reason why the validity of all tests was and continues to be an evolving property and validation is an ongoing process.

To validate an interpretative inference is to determine the degree to which multiple sources of evidence are consistent with the inference, while establishing that alternative inferences are less well supported. To validate an inference about the implications for action requires validation not only of score meaning, but also of value implications and action outcomes, particularly appraisals of the relevance and utility of the test scores for specific applied purposes and of the social consequences of using the scores for applied decision making (Messick, 1989). Therefore, the key issues of test validity are the interpretability, relevance and utility of scores, the value implications of scores as a basis for action, and the functional worth of the scores in terms of the social consequences of their use (Messick, 1994).

For example, an occupational therapist completes the *Kohlman Evaluation of Living Skills* (KELS) with a 76-year-old male client who sustained a recent stroke who was admitted to a rehabilitation facility 4 weeks ago to assess his instrumental (IADL) and basic activities of daily living (ADL) performance skills. The KELS score will be used to in part to determine if the treatment team working with the client and his family will make the recommendation whether the client can be discharged home safely or whether he should be discharged to a skilled nursing facility. The functional worth of the KELS score has many potential social consequences for the client's living arrangements, health and well-being. The therapist using the KELS needs to be sure that it is valid since the recommendations made based on KELS's scores directly impact the lives of individuals.

It is important to note, however, that validity is a matter of degree rather than all or none. Over time, the existing validity becomes enhanced or contravened by new research findings, and projections about the potential social consequences of testing become transformed by evidence of actual consequences and by changing social conditions. In this way, validity is an evolving property and validation is an ongoing dynamic process. Since the validity evidence of an instrument is always incomplete, validation is essentially a matter of making the most reasonable case to guide both current use of an instrument and current research to advance understanding of what the instrument scores mean. Hence, validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on instrument scores (e.g. recommendations that are made based clients' performance scores on the KELS). It is important, therefore, for occupational therapists to stay up-to-date with the current body of validity evidence related to the tests, instruments, and scales they use in their daily clinical practice.

Construct Validity: A Unitary View of Validity

In 1989, Messick proposed that all components of validity be subsumed under the concept of construct validity, a claim that is now gaining wider acceptance in the research, testing, education, measurement and professional communities (AERA, APA, & NCME, 1999). He proposed an overall unifying concept of validity that takes into account both score meaning and social values in test/scale interpretation and test/scale use. This current-day conceptualization of validity, known as "construct validity," integrates the traditional components of content, criterion-related, and construct validity.

In the 1999 *Standards* (AERA, APA, & NCME, 1999), five sub-components of construct validity evidence are included as a means of addressing the central issues implicit in the notion of validity as a unified concept. These subcomponents are (a) test content evidence, (b) response processes evidence, (c) internal structure evidence (d) relations to other variables evidence, and (e) consequences of testing (see Figure 2). They function either as general validity criteria or as standards for all measurement. Taken together, the five components provide a way of addressing the multiple and interrelated validity questions that need to be answered in order to justify test score interpretation and use. The five sources of construct validity evidence as reported in the *Standards* (AERA, APA, & NCME, 1999) are listed in Table 1 and are outlined for the reader below.

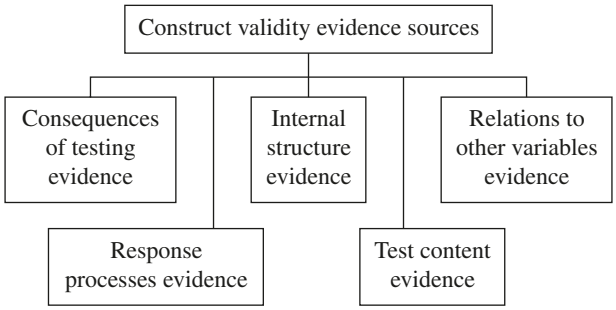

Figure 2. *Sources of construct validity evidence (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).*

1. Test Content Evidence

An essential issue for the content aspect of construct validity evidence is the delineation of the boundaries of the construct domain being assessed. Take the situation when an occupational therapist gets a referral from a pediatrician requesting that she assess the social skills of an 8-year-old boy with Asperger's Syndrome. The assessment results are used to determine whether the boy will meet the eligibility criteria to be enrolled in a social skills program for students run at a local community health centre. The therapist would need to ensure that any test or scale she selected to assess the boy would need to include items and subscales that were reflective of the abilities and skills (in this case social skills) she was requested to assess. In other words, any social skills scale she chose to use would need adequate *test content evidence*.

The test content evidence aspect answers the question: "To what extent does the content of the test, including items and formats, adequately reflect the content domain?" "Traditional considerations of content treat the test as a sample from some specified behavioral domain or item universe about which inferences are to be drawn or predictions made" (Messick, 1989, p. 36). However, it is not enough to merely select tasks/test items that are relevant to the construct domain. The items should include tasks that are representative of the domain so as to ensure that all the important parts of the construct domain are covered (e.g. determining the knowledge, skills, attitudes, motives, and other attributes to be specified by the assessment tasks).

This type of validity evidence is based on logical analyses and evaluations of the test content by experts including items, tasks formats, wording, and demands placed on respondents completing the test items/assessment tasks. "In general, it addresses questions about the extent to which the content of a measure represents a specified content domain" (Goodwin & Leech, 2003, p. 183), as well as the relevance of the construct domain to the proposed interpretations of scores obtained with the test. Expert evaluations and reviews are completed to generate evidence about a test's features including sufficiency, clarity, relevancy, comprehensiveness, and the commonality

Table 1. Validity evidence types and examples of validation activities

Type of validity in 1999 *Standards*[a]	Type of validity in 1985 *Standards*[b]	Question the type of evidence attempts to answer	Examples of validation activities
1. Test content evidence: test items should include tasks that are representative of the domain so as to ensure that all the important parts of the construct domain are covered (e.g. determining the knowledge, skills, attitudes, motives, and other attributes to be specified by the assessment tasks)	Construct validity evidence	"To what extent does the content of the test, including items, subscales, and formats, adequately and comprehensively represent the content domain?"	• Test blueprint and specifications • Representativeness of the test blueprint of the construct domain • Test item writers' qualifications, experience, knowledge, and expertise • Sensitivity review • Item technical quality • Logical analyses and experts' evaluations of the degree that the test content represents the content domain • Logical analyses and experts' evaluations to the degree that the items, assessment tasks, or subscales of a test fit the definition of the construct or test purpose • Logical analyses and experts' evaluation of the relevance, importance, clarity, and lack of bias in a test's items, subscales, or assessment tasks • Logical analyses and experts' opinions of the degree that construct underrepresentation or construct-irrelevant aspects of a test may result in disadvantages for one or more subgroup(s) of respondents and unfair advantages for one or more sub group of respondent(s).
2. Response processes evidence: evaluates the extent to which tasks/items on a test or the type of responses required of respondents fit the intended, defined construct	Construct validity evidence	"To what extent does the type of performances or responses of the individuals completing the test fit the intended construct being measured or evaluated?"	• Evaluation of participants' responses to test items and performance tasks using participant debriefing interviews • Investigations of the ways that raters, observers, interviewers, and judges collect and interpret data • Longitudinal investigations of changes or trends in patterns of respondents' answers to items or tasks • Respondents' familiarity with test format • Respondents' accuracy when different answer format scores are combined • Quality control/accuracy of final scores and grades • Quality control of score reporting to respondents and examiners • Understandable and accurate interpretations and reporting of scores to respondents • Behavioral observation • Person Fit & Item Difficulty Hierarchy (from Rasch Measurement Model data analysis output)
3. Internal structure evidence: the statistical or psychometric traits of the test items, the test properties (such as reproducibility and generalizability), and the measurement model used to score and scale the test	Construct validity evidence	"To what extent does the relationships between the test items match the construct (as it is defined) being measured or evaluated?"	• Exploratory and confirmatory factor analysis of scale items • Evaluating dimensionality of scale items • Cluster analysis of scale items • Test score reliability • Subscale correlations • Evaluation of scale item relationships using item analysis techniques including: item difficulty/ discrimination, item/test characteristic curves, interitem correlations and item-total correlations • Completion of studies that evaluate differential item functioning of items • Examination of hierarchical ordering of scale items based on item difficulty

(*Contd*)

Table 1. (*Continued*)

Type of validity in 1999 *Standards*[a]	Type of validity in 1985 *Standards*[b]	Question the type of evidence attempts to answer	Examples of validation activities
4. Relations to other variables evidence: evaluates the extent to which test score properties and interpretations can be generalized to and across sample groups, settings, and tasks. It includes validity generalization of test criterion relationships based on relations to other variables	Criterion-related evidence (concurrent and predictive validity)	"What is the type and extent of the relationships between test scores and other variables such as those the test is expected to correlate with or to predict?	• Correlation studies of the relationships between test scores and external criterion variables (concurrent validity) • Correlation studies of the extent to which scores obtained on the test in question predict external criterion variables evaluated at a later date (predictive validity) • Differential group relationship or prediction studies • Investigations of the effectiveness and accuracy of test scores in selection, classification, and placement decisions • Validity generalization studies
	Construct validity (convergent, divergent, and discriminant validity)	"What is the type and extent of the relationships between test scores and other variables such as those the test is expected to correlate with?" and "Is the test able to differentiate between two groups of test takers/respondents with a known difference? (e.g. age, gender, diagnosis)?"	• Convergent validity studies that evaluate the correlational relationship of test scores and other similar variables that the test scores should theoretically have high correlations with • Divergent validity studies that evaluate the correlational relationship of test scores and other dissimilar variables that test scores should theoretically have low or no correlations with • Experimental studies that evaluate hypotheses about intervention effects on scores obtained with a scale • Known group comparison investigations that evaluate hypotheses about expected differences in mean scores between groups of respondents (also known as discriminant validity) • Longitudinal studies that test hypotheses about expected differences between two groups with known differences in mean test scores over time
5. Consequences of testing evidence: evaluates the value implications of score interpretations as a basis for action, as well as the actual and potential consequences (anticipated and unanticipated consequences) of test use in the short-term and long-term, especially in regards to sources of invalidity related to issues of bias, fairness, equity, and distributive justice		"To what extent are the anticipated benefits of testing realized?" and "To what extent do unanticipated benefits, both negative and positive, occur?"	• Investigations of the degree of unexpected or unanticipated negative consequences of testing happen • Investigations of the degree of expected or anticipated benefits of testing are realized • Analysis of impact of test scores/results on respondents and society at large • Evaluation of the consequences of testing on students' future learning • Analysis of false positives and false negatives test results • Reasonableness of method used to establish pass/fail (cut) score

Sources: Goodwin, 2002b; Goodwin & Leech, 2003; Downing, 2003; Wolfe & Smith, 2007a, 2007b. [a]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. [b]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.

between the items and tasks and the definition of the construct being measured.

Bias (such as gender, age, culture, socioeconomic) is another focus of the review process. Included in this arena are the notions of "construct-irrelevant variance" and "construct underrepresentation" which are related to the extent to which the test appears to measure more or less of the construct than what is intended. Construct-irrelevant variance (also known as surplus construct irrelevancy), which occurs when an assessment is too broad, "containing excess reliable variance associated with other distinct constructs as well as method variance such as response in a matter irrelevant to the interpreted construct" (Messick, 1995, p. 742). In other words, the test contains excess reliable variance that is irrelevant to the interpreted construct. Therefore, a primary validation concern is the extent to which one scale might underrepresent the construct it purports to measure while concurrently contaminating the scale scores with construct-irrelevant variance. Construct underrepresentation occurs when a test is too narrow or fails to include important features, components, dimensions, or facets of the construct in question.

To avoid either construct under-representation or construct-irrelevant variance, Messick (1989) suggested that the content specifications of a scale reflect the breadth and scope of the construct invoked in score interpretation. For example, self-esteem is composed of a number of subconstructs. If each of these sub-constructs are considered on their own, each component would under-represent some aspects of the overall construct; however, a composite of all of the relevant self-esteem sub-constructs would reduce the chances of construct under-representation from occurring. Similarly, Messick (1989) suggested that incorporating multiple item or task formats in the total-score composite of a construct is another way to ensure that construct under-representation does not occur.

Traditionally, content relevance and representativeness of assessment tasks have been appraised by expert professional judgment. Documentation of the input from the panel of knowledge experts serves as a means to address the content aspect of construct validity (Streiner & Norman, 1995). However, judgments of the relevance of test items or tasks to the intended score interpretation need to take into account all aspects of the testing procedure that significantly affect the performance of test respondents. These include specification of the construct domain of reference in terms of topical content, typical behaviors, and underlying processes. Test specifications regarding stimulus formats and response alternatives, administrative conditions (such as instructions of time limits for respondents completing the test), and criteria for item scoring are other factors that need consideration (Anastasi & Urbina, 1997).

Other sources of content evidence are the plan and outline of the test in the form of a detailed test blueprint (also known as test specifications) that relates to the content being assessed. The test blueprint needs to be detailed enough to outline the subclassifications of content and specify the percentage of test questions per content category and the difficulty level of the test questions. The test blueprint must also be directly related to the educational objectives or body of knowledge they are intended to evaluate. Independent expert review of the test blueprint can provide objective feedback about the test items in relation to learning objectives evaluated and what difficulty level (Downing, 2003).

2. Response Processes Evidence

In Messick's framework, this component was referred to as the substantive aspect of construct validity. In the 1985 edition of the *Standards* (AERA, APA, & NCME), this type of evidence (based on response processes) was included as a component of construct validity evidence. The response processes source of validity evidence answers the question: "To what extent does the type of performances or responses of the individuals completing the test fit the intended construct?" Evidence based on response processes evaluates the extent to which tasks/items on a test or the type of responses required of respondents fit the intended, defined construct. For example, a test of fine motor skills does not include test items related to static and dynamic balance skills, or jumping abilities, but rather would likely include items related to manual dexterity, in-hand manipulation skills, eye-hand coordination, pencil grasp, grip strength, and/or visual-motor integration skills.

The response process evidence aspect of construct validity refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks/items. It emphasizes the role of substantive theories and process modeling in identifying the domain processes to be revealed in assessment tasks/items (Messick, 1989). For example, items on a sensory integration dysfunction screening scale are based on sensory integration theory (Bundy, Lane, & Murray, 1991).

In the response process approach, assessment tasks and/or items are included in the original item pool on the basis of judged relevance to a broadly defined domain, but they are selected for the specific test on the basis of empirical response consistencies. Two significant points are involved. The first is the need for tasks/items that will provide appropriate sampling of domain processes in addition to the traditional coverage of domain content. The second point is the need to

move beyond traditional expert opinion of content to the accumulation of scientific evidence that demonstrates whether the ostensibly sampled processes are truly engaged by the respondents completing the assessment tasks and answering assessment items (Messick, 1989). For example, if the items on a sensory integration screening scale are too difficult or not engaging for the respondents completing them even though the scale items represent the theoretical concepts, then the scale items are problematic.

In the 1999 *Standards* context, response process was defined as evidence of data integrity such that all sources of potential error associated with test administration are controlled or eliminated to the largest extent possible. Response process has to do with all components of testing such as ensuring the accuracy of all responses to test prompts, the quality control of data coming from tests, the suitability of the methods used to combine different types of test scores into one composite score, and the usefulness and accuracy of the score reports provided to respondents (Downing, 2003). Documentation of all quality-control procedures used to ensure high level accuracy of test scores is also an integral source of evidence.

Other sources of response processes can include documentation of the reasons for the types of scores reported, the method used to report test scores, the explanation and interpretation materials provided to explain the score report, and the inclusion of materials that discuss the proper use of and common misuses of the test score data. Often results from commonly used tests such as the *Developmental Profile III, Peabody Developmental Motor Scales, 2nd edition*, *Miller Function and Participation Scales*, or the BOT-2 include an accompanying information sheet explaining the meaning of scores provided to parents of children.

Ways to gather evidence based on response processes include observing respondents completing required test tasks or interviewing respondents to determine reasons why he/she provided certain answers to questions. Another source of response process evidence is to investigate the ways in which observers, judges, and raters use criteria to record and score respondents' behavior, respondents' performance on assessment tasks or essays/assignments completed by respondents. Some tests actually provide information about raters' leniency/severity when scoring items (e.g. *Test of Playfulness*, *Assessment of Motor and Process Skills, Evaluation of Social Interaction*). This information can be used to provide valuable information under the response processes evidence category. What is being investigated here is whether raters are using the scoring criteria as intended or whether they are using irrelevant or extraneous factors that do not fall within the planned interpretation of scores.

3. Internal Structure Evidence

The internal structure aspect of construct validity evaluates the fidelity of the scoring structure to the construct domain structure. In the 1985 *Standards,* this type of evidence was considered part of the construct validity evidence. "It examined the extent to which the internal components of a test match the defined construct and is most often estimated by confirmatory factor analysis" (Goodwin & Leech, 2003, p. 184). In the 1999 *Standards*, internal structure referred to the statistical or psychometric traits of the test items, the test properties (such as reproducibility and generalizability), and the measurement model used to score and scale the test. The internal structure aspect of validity evidence answers the question: "To what extent do the relationships between the test items match the construct (as it is defined)?"

It should be noted that the internal structure aspect of validity now includes what was traditionally viewed as the separate category of reliability (e.g. intra-rater, inter-rater, test-retest, internal consistency, alternate form and split-half). In other words, the reliability of an instrument now provides evidence about its "internal structure."

Scoring models should be consistent with what is known about the structural relations inherent in the behavioral components of the construct in question. In ideal circumstances, the manner in which behavioral components are joined together to produce a score should be dependent on the knowledge of how the processes underlying those behaviors combine dynamically to create effects (Messick, 1995). For example, in the *Assessment of Motor and Process Skills* (AMPS) is an observational assessment that is used to measure the quality of a person's activities of daily living (ADL). The AMPS is completed by a therapist rating the quality of a person's ADL performance based on his/her effort, efficiency, safety, and independence on 16 ADL motor and 20 ADL process skill items. Examples of the motor skill items include "walk, reach, lift, align, grip, and transport" and process skills include "search, gather, organize, restore, and navigate." The authors of the AMPS have determined that the 16 ADL motor and 20 ADL process skill items are representative enough of the ADL construct it purposes to measure. The internal structure of the scale (such as the AMPS or BOT-2) should be consistent with what is known about the internal structure of the construct domain being assessed. Basing the internal structure of the items of a scale or instrument on what is known about the internal structure of the construct domain being measured is called *structural fidelity* (Messick, 1995).

Differential item functioning (DIF) has also been suggested as another category of evidence for internal structure. DIF refers to the situation where respondents of equal ability perform

differently on a test due to items that are biased against groups based on such variables as age, gender, or ethnicity. Therefore, DIF studies are completed to detect test item bias.

Many of the statistical analyses completed to support or refute evidence of a test's internal structure are often done as routine quality-control procedures (Downing, 2003). Examples of this are item analyses that compute the difficulty/easiness level of each test item, the discrimination of each test item (a statistical index indicating how well a test item separates high scoring from low scoring respondents) and a count of the proportion of respondents who completed each option to a test question. Summary statistics profiling a whole test can be calculated and examples of these include overall test difficulty/easiness, the average test discrimination and the internal consistency reliability of the test.

Different types of reliability such as test-retest reliability, inter-rater reliability, and intra-rater reliability all contribute to the internal structure validity evidence base of an instrument. For example, on a test that assesses school-age children's visual motor integration skills, their abilities to copy geometric shapes and designs based on the quality of the drawing and copying output are rated. The intra- and inter-rater reliability of the visual motor integration test would need to be evaluated. Issues of bias, sensitivity and fairness also relate to the internal structure of a test and are important sources of validity evidence (Downing, 2003).

4. Relationships to Other Variables Evidence

In the construct validity framework put forth by Messick, the "relationships to other variables" evidence category was referred to as generalizability evidence and external evidence. This source of validity evidence answers the question: "What is the type and extent of the relationships between test scores and other variables such as those the test is expected to correlate with or to predict?" The generalizability aspect of construct validity evaluates the extent to which test score properties and interpretations can be generalized to and across sample groups, settings, and tasks. It includes validity generalization of test criterion relationships based on relations to other variables (Messick, 1989). The degree of generality of construct meaning across contexts may be appraised by all of the techniques of construct validation. These include assessing the extent to which test scores reflect comparable patterns of relationships with other measures, common underlying processes, or similar responsiveness to treatment across groups, situations, times, and tasks. The generality component of construct validity attempts to ensure that a test provides representative coverage of the content and processes of the construct domain in question. In other words, a test that evaluates a participant's level of consciousness, such as the *Glasgow Coma Scale*, adequately covers all the components that are included in the construct of consciousness. It is meant to ensure that the score interpretation is not limited to the sample of assessed tasks, but is also broadly generalizable to the construct domain.

Evidence of such generalizability depends on the degree to which assessment tasks can be correlated with other tasks representing the same construct or components of the construct. For example, the *Developmental Test of Visual Motor Integration,* 5th edition, *Full Range Test of Visual Motor Integration*, *Test of Visual-Motor Skills – Revised*, and the *Slosson Test of Visual Motor Performance* all state that they measure the construct of visual-motor integration. Hence, if performance data on these four tests were collected from the same group of respondents, then one would expect the test scores to be highly correlated with one another if in fact they did measure the same visual-motor integration ability construct. Since generalizability evidence helps to establish the boundaries of score meaning, it is important to ensure that the range of items and assessment tasks included in a scale or instrument represent the full range of the construct being measured. The "relationship to other variables" evidence category is broad and includes several traditional types of validity including criterion-related, concurrent, predictive, convergent, and discriminant validity. Commonly used approaches to collect this type of validity evidence include experimental studies, correlation studies, statistical investigations, and known-group/criterion-group comparison studies.

Evidence of relationships to other variables "refers to the extent to which the assessment scores' relationships with other measures and non-assessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed" (Messick, 1995, p. 746). The meaning of scale scores is externally substantiated by evaluating the degree to which empirical relationships with other scale scores or the lack thereof are consistent with that meaning. In other words, the constructs represented in the test should rationally account for the external pattern of test correlations. Convergent, divergent, and discriminant correlation patterns with external variables are significant because the convergent pattern indicates a correspondence between scale measures of the same construct, and the divergent/discriminant pattern indicates a distinctness from measures of other unrelated constructs. Convergent, divergent, and discriminant evidence are basic to the construct validation process. Traditionally, in the tripartite validity context, these types of validity were associated with construct validity.

The concept of convergence and divergence of validity evidence is demonstrated by the research design proposed by Campbell and Fiske (1959) referred to as the "multitrait-multimethod approach." In this design, different tests of the

same construct (achievement, ability, performance) are correlated with different tests of the same construct. The resulting pattern of correlation scores may demonstrate the convergence or divergence of the different assessment methods on tests of the same and different abilities or skills.

Since typical task performance takes a great deal of time, there is often a conflict in assessments of the time-intensive depth of examination and the breadth of domain coverage needed for generalizability of construct interpretation. The conflict between the assessment task completion time and the degree of assessment task detail to be included is frequently viewed as a trade-off between validity and reliability of generalization. "It might better be depicted as a tradeoff between the valid description of the specifics of a complex task and the power of construct interpretation" (Messick, 1995, p. 746).

In addition to generalizability across assessment tasks and scale items, the limits of score meaning are also affected by the degree of generalizability across time and occasions and across observers or raters of assessment task performance (Messick, 1995). An example of the trade-off between assessment task completion time and the detail of assessment task items is the BOT-2 and the *Movement Assessment Battery for Children, 2nd edition* (MABC-2). The BOT-2 is a detailed motor skill test that includes eight sub-scales comprised of over 50 items whereas the MABC-2 is a screening tool that includes three fine motor assessment tasks and six gross motor items. Another example would be the long and short forms of the BOT-2 or the Sensory Profile. Decisions have to be made about the breadth and depth of detail and scope of a test that are needed in clinical contexts. An occupational therapist's clinical reasoning skills, professional knowledge, and contextual awareness would help inform what breadth and depth of information is required.

5. Consequences of Testing Evidence

The "consequences of testing" aspect of construct validity evaluates the value implications of score interpretations as a basis for action, as well as the actual and potential consequences (anticipated and unanticipated consequences) of test use in the short-term and long-term, especially in regards to sources of invalidity related to issues of bias, fairness, equity, and distributive justice. In other words, the social consequences of testing may be either positive or negative when associated with bias in scoring and interpretation or with unfairness in test use. In this situation, the primary concern with respect to adverse consequences is that any negative impact on individuals or groups should not be caused by any source of test invalidity such as construct under-representation, construct-irrelevant variance, differential item functioning, or test item bias (Messick, 1995).

This aspect of validity was virtually absent from the 1985 *Standards*. This is significant since consequential evidence is concerned with the potential implications or impact, both positive and negative, that test use can have on respondents. For example, certain items on a test/scale may discriminate against certain respondents due to gender, language, ethnicity, cognitive impairment, socioeconomic level, age, or physical disability. The consequential category of construct validity evidence addresses the questions: "To what extent are the anticipated benefits of testing realized?" and "to what extent do unanticipated benefits, both negative and positive, occur?" "Evidence related to consequences of testing and its outcomes is presented to suggest that no harm comes directly from the assessment or, at the very least, more good than harm arises from the assessment" (Downing, 2003, p. 836).

An example of the "consequences of testing evidence" related to occupational therapy practice would be the use of the *Functional Independence Measure* (FIM) in rehabilitation settings. Take the case of a 20 year old man has been admitted to an inpatient rehabilitation centre with the diagnosis of a traumatic brain injury after being medically stabilized at an acute care hospital. His health insurance company has agreed to pay for two weeks of intensive rehabilitation, but beyond that, funding for his treatment is dependent on the functional gains he has made as indicated by his weekly FIM score. Often the FIM is used as an outcome measure benchmark by health insurance agencies to monitor a patient's progress. If the patient's FIM score does not indicate sufficient functional change, then the patient's health insurance benefits to cover the cost of his ongoing rehabilitation may be terminated. This would have potentially negative consequences for the patient's ongoing recovery and long term prognosis.

High-stake examinations such as the United States Medical Licensure Examination sequence (sponsored by the National Board of Medical Examiners) and the Occupational Therapy Registration examination (sponsored by the National Board for Certification in Occupational Therapy, Inc.) are examples of summative professional hurdle requirements in order to practice in a specific profession. If respondents do not pass these examinations, the potential negative consequences are great. Therefore, the issue of false positives (candidates who pass an exam who should have actually failed) and false negatives (candidates who fail an exam who should have actually passed) in such exams may cause harm to the clients they serve. In this context, respondents failing either these examinations can be seen as both a positive and negative consequence of testing. The negative social consequence is for the student, who after completing a long, expensive, and challenging professional training program, is not allowed to practice clinically.

The positive social consequence is that hopefully the general public is protected from a poorly skilled clinician being qualified and then potentially causing harm to clients.

Consequential validity evidence is also related to the issue of establishing pass rates for tests (also known as cut scores), the statistical properties of passing scores, documentation of the method used to establish pass-fail score and the rationale for the selection of a specific passing score method. Other psychometric indicators about the passing score and its consequences include a formal, statistical estimation of the pass-fail decision reliability or classification accuracy and some form of estimation of the standard error of measurement at the cut score. Given the regency of this type of validity, there has been little consideration of the consequences of testing in the literature (AERA, APA, & NCME, 1999). Chudowsky and Behuniak (1998) suggested that using focus groups as a means to examine the consequential validity of a large scale assessment program.

Conclusion

The meaning of validity has evolved with the new edition of the *Standards* (AERA, APA, & NCME, 1999) containing a revised definition and description of validity that eliminates the content, criterion-related, and construct types of validity. The contemporary view of validity is seen as a unitary concept known as construct validity. The traditional and popular "tripartite" or "trinity" view of validity has been replaced by one that promotes five distinct types of validity evidence based on content, response processes, internal structure, relations to other variables, and consequences of testing. The tripartite validity view was problematic for a number of reasons, including (a) it masked the unitary nature of validity, (b) it compartmentalized how validity was conceptualized in narrow categories, (c) it was incomplete since it ignored the testing consequences question, and (d) it promoted the idea that all types of validity were equal. "Approaching validity as multidimensional and complex—requiring a wide and diverse body of evidence—is much more realistic and appropriate" (Goodwin & Leech, 2003, p. 189).

Hence, it is essential that occupational therapists are knowledgeable about construct validity as a unitary concept due to its impact on evidence-based practice, validation of occupational therapy based practice models, generation of psychometrically sound tests and scales, and the use of valid tests to substantiate treatment planning and intervention provision. If therapists use tests, instruments, and scales (that do not have well established construct validity evidence) to generate research data and treatment outcome findings, that are used to make claims about the effectiveness of occupational therapy intervention, then the conclusions drawn from those studies will questionable and not sound. Hence, construct validity has a direct impact on the evidence-based practice of occupational therapy. There are ongoing issues and challenges that face the validity debate and occupational therapists need to be conversant with them.

References

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

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

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

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

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

Anastasi, A. (1988). *Psychological testing*. New York: Macmillan Publishing.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Upper Saddle River, NJ: Prentice Hall.

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

Asher, I. E. (2007). *Occupational therapy assessment: An annotated index*. Bethesda, MD: American Occupational Therapy Association.

Benson, J., & Schnell, B. A. (1997). Measurement theory: Application to occupational and physical therapy. In: J. Van Deusen & D. Brunt (Eds.), *Assessment in occupational therapy and physical therapy* (pp. 3–24). Philadelphia, PA: W. B. Saunders.

Bundy, A. C., Lane, S. L., & Murray, E. A. (1991). *Sensory integration: Theory and practice*. Philadelphia, PA: F. A. Davis.

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 aspects of validity. *Educational Measurement: Issues and Practice, 17*(4), 28–38.

Craik, J., Davis, J., & Polatajko, H. (2007). Introducing the Canadian Practice Process Framework. In: E. A. Townsend & H. J. Polatajko (Eds.), *Enabling occupation II: Advancing an occupational therapy vision for health, well-being, & justice through occupation* (pp. 229–246). Ottawa, ON: Canadian Association of Occupational Therapists.

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In: H. Wainer & H. I. Baum (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, IL: University of Illinois Press.

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

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*, 830–837.

Downing, S. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*, 327.

Fawcett, A. L. (2007). *Principles of assessment and outcome measurement for occupational therapists and physiotherapists: Theory, skills and application*. Hoboken, NJ: Wiley.

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*(3), 102–107.

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 & Nutrition, 35*(1)*,* 6–7.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the *New Standards for Educational and Psychological Testing*: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181–191.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36,* 456–462.

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

Hinojosa, J., & Kramer, P. (1998). *Evaluation: Obtaining and interpreting data*. Bethesda, MD: American Occupational Therapy Association.

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*(5), 736–753.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 11*, 527–535.

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation & the Health Professions, 17*(2), 133–159.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

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

Kielhofner, G. (2006). Developing and evaluating quantitative data collection instruments. In: G. Kielhofner (Ed.), *Research in occupational therapy: Methods of inquiry for enhancing practice* (pp. 155–176). Philadelphia, PA: F. A. Davis Company.

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.

Law, M., & Baum, C. (2005). Measurement in occupational therapy. In: M. Law, C. Baum, & W. Dunn (Eds.) *Measuring occupational performance: Supporting best practice in occupational therapy* (pp. 3–20). Thorofare, NJ: Slack.

Law, M., Baum, C., & Dunn, W. (2005). *Measuring occupational performance: Supporting best practice in occupational therapy*. Thorofare, NJ: Slack.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36,* 437–448.

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

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: Lawrence Erlbaum.

Messick, S. (1989). Validity. In: R. Linn (Ed.), *Educational measurement* (pp. 13–104). New York: American Council on Education & Macmillan Publishing.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36,* 463–469.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory*. New York: McGraw-Hill.

Portney, L., & Watkins, M. (2000). *Foundations of clinical research.* Sydney, NSW: Prentice-Hall.

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

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36,* 477–481.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*, 281–311.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use*. Oxford, UK: Oxford University Press.

Wolfe, E. W., & Smith, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I – instrument development tools. *Journal of Applied Measurement, 8*(1), 97–123.

Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II – validation activities. *Journal of Applied Measurement, 8*(2), 204 – 234.

Yun, J., & Ulrich, D. A. (2002). Estimating measurement validity: A tutorial. *Adapted Physical Activity Quarterly, 19*, 32–47.