

## Chapter 2

# Standards and Guidelines for Validation Practices: Development and Evaluation of Measurement Instruments

Eric K.H. Chan

This book, *Validity and Validation in Social, Behavioral, and Health Sciences* (edited by Zumbo and Chan), is a collection of chapters synthesizing the practices of measurement validation across a number of academic disciplines. The objectives of this chapter are to provide an overview of standards and guidelines relevant to the development and evaluation of measurement instruments in education, psychology, business, and health. Specifically, this chapter focuses on (1) reviewing standards and guidelines for validation practices adopted by major professional associations and organizations and (2) examining the extent to which these standards and guidelines reflect contemporary views of validity, and issues, topics, and foci considered therein (e.g., Kane 2006, 2013; Messick 1989; Zumbo 2007, 2009).

### Validity and Validation

Measurement instruments are widely used for clinical, research, and policy decision making purposes in many professional disciplines. The quality of the data (i.e., reliability) and the quality of the decisions and inferences made based on the scores from measurement instruments (i.e., validity) are therefore not inconsequential. Validity and validation are the most fundamental issues in the development, evaluation, and use of measurement instruments. *Validity* refers to the quality of the inferences, claims, or decisions drawn from the scores of an instrument and *validation* is the process in which we gather and evaluate the evidence to support the appropriateness, meaningfulness, and usefulness of the decisions and inferences

---

E.K.H. Chan (✉)

Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada  
e-mail: [eric.chan.phd@gmail.com](mailto:eric.chan.phd@gmail.com)

that can be made from instrument scores (i.e., to understand and support the properties of an instrument) (Zumbo 2007, 2009).

Although it is not unanimous (see, for example, Borsboom et al. 2004; Markus and Borsboom 2013 as dissenting views), overall there are a series of statements about validity and validation practices that are shared and characterize a “contemporary view of validity” (e.g., Cronbach 1988; Hubley and Zumbo 1996, 2011, 2013; Kane 2006, 2013; Messick 1989; Zumbo 2007, 2009):

1. Validity is about the inferences, claims, or decisions that we make based on instrument scores, not the instrument itself.
2. Construct validity is the focus of validity. Validity does not exist as distinct types and validation should not be a piecemeal activity. Sources of validity evidence are accumulated and synthesized to support the construct validity of the interpretation and use of instruments.
3. Validation is an ongoing process in which we accumulate and synthesize validity evidence to support the inferences, interpretations, claims, actions, or decisions we make.
4. The contemporary views of validity contend that in addition to the traditional sources of validity such as content, relations to other variables (e.g., convergent, discriminant, concurrent, and predictive validity), and internal structure (dimensionality), evidence based on response processes (cognitive processes during item responding or during rating) and consequences (the intended use and misuse) are important sources of validity evidence that should be included in validation practices. These sources of evidence are accumulated and synthesized to support the validity of score interpretations.
5. Although different validity theorists emphasize each of these to varying amounts, validation practices center around establishing a validity argument (Cronbach and Kane), an explanation for score variation (Zumbo), the substantive aspect of construct validity, which highlights the importance of theories and process modeling that are involved in item responses (Messick), sample heterogeneity and exchangeability to support inferences (Zumbo), or being guided by a progressive matrix that organizes validation practices, but centers on construct validity (Messick).

## Standards and Guidelines

Standards and guidelines play an important role in professional practices. They make professional practices more efficient and consistent, bridge the gap between what the empirical evidence supports and what professionals do in practice, and serve as gatekeepers to ensure high quality professional practice (Woolf et al. 1999). Although it is not the intent of this chapter to discuss the differences between standards and guidelines, it is worth noting that the two are not the same. According to the American Psychological Association (APA 2002a).

The term *guidelines* [italics in original] refers to pronouncements, statements, or declarations that suggest or recommend specific professional behavior, endeavors, or conduct . . . Guidelines differ from standards in that standards are mandatory and may be accompanied by an enforcement mechanism. Thus . . . guidelines are aspirational in intent. They are intended to facilitate the continued systematic development of the profession and to help ensure a high level of professional practice . . . Guidelines are not intended to be mandatory or exhaustive and may not be applicable to every professional and . . . [professional] situation. They are not definitive and they are not intended to take precedence over [professional judgment]. (p. 1050)

Guidelines on the development of guidelines are available (APA 2002a; Eccles et al. 2012; Shekelle et al. 1999), as are criteria for evaluating the quality of guidelines (APA 2002b; The AGREE Collaboration 2003). Over the years standards and guidelines have been developed by a number of organizations in various disciplines (including education, health, medicine, and psychology) regarding the development and evaluation of measurement instruments. It is important to note that the purpose of this chapter is not on the quality appraisal of the standards and guidelines, but rather on informing the readers on the issues of validity and validation as covered in the standards and guidelines, as well as on examining the extent to which the standards and guidelines reflect contemporary views of validity. In this chapter, the following standards and guidelines are covered:

1. *Standards for Educational and Psychological Testing* (AERA et al. 1999)<sup>1</sup>
2. *Guidance for Industry – Patient-Reported Outcomes Measures: Use in Medical Product Development to Support Labeling Claims* (Food and Drug Administration 2009)<sup>2</sup>
3. *Consensus-Based Standards for the Selection of Health Measurement Instruments* (COSMIN; Mokkink et al. 2010a)
4. *Evaluating the Measurement of Patient-Reported Outcomes* (EMPRO; Valderas et al. 2008)
5. *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology 2003)
6. Test Reviewing for the *Mental Measurement Yearbook* at the Buros Center for Testing (Carlson and Geisinger 2012)
7. European Federation of Psychologists' Association's (EFPA) review model (Evers et al. 2013)

---

<sup>1</sup> The International Test Commission (ITC 2001) has guidelines on test use. Although the guidelines, as stated in the document, have implications on the development of measurement instruments, the focus is on test user competencies (e.g., knowledge, skills, abilities, and related characteristics). The ITC guidelines are therefore not included in this review.

<sup>2</sup> The European Medicines Agency (EMA 2005) published a document providing broad recommendations on the use of health-related quality of life (HRQoL), a specific type of patient-reported outcomes (PRO), in their medical product evaluation process. The EMA explicitly states that it is a reflection paper, *not* guidance. Therefore, the EMA document is not included in the present review.

## Standards for Educational and Psychological Testing

The development of the *Test Standards* began when the APA published a formal proposal (*Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal*) in 1952 on the standards to be used in the development, use, and interpretation of measurement psychological instruments. The proposal led to the publication of the first standards in 1954, the *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. In the document, validity was classified into content, predictive, concurrent, and construct. The *Test Standards* have undergone several revisions (APA 1966; AERA et al. 1974, 1985). The most current version of the *Test Standards* (AERA et al. 1999) is clearly heavily influenced by Messick's (1989) unitary view of validity. Accordingly, there is no singular source of evidence sufficient to support a validity claim. Construct validity is the central component in validation work, encompasses the following five sources of evidence germane to the validation of the interpretation and use of the score of an instrument. The five sources include (1) evidence based on test content, (2) evidence based on response processes, (3) evidence based on internal structure, (4) evidence based on relations to other variables, and (5) consequences. A cursory review of the forthcoming edition of the *Test Standards* suggests that, overall, the focus and orientation of the 1999 edition are maintained.

The content of an instrument includes the items, format and wording of the items, response options, and the administration and scoring procedures. Content evidence can be obtained by examining the relationship between the content of an instrument and the construct one intends to measure. Evidence based on response processes is the examination of the cognitive or thinking processes involved when people respond to items. Strategies such as think aloud protocols can be used to investigate how people interpret and answer items. The internal structure of an instrument refers to the degree to which the items represent the construct of interest by investigating how items relate to each other using statistical methods such as factor analysis and item response modeling. Evidence based on relations to other variables concerns the association between instrument scores and external variables. Convergent, discriminant, and criterion-related (including concurrent and predictive) validity can be gathered to support such evidence. And finally, consequences refer to the intended and unintended use of an instrument and how its unintended use weakens score inferences. Table 2.1 presents the sources of evidence discussed in the *Test Standards*.

It is noteworthy that the APA, which publishes the *Test Standards*, appears to be using the term "standards" in a manner inconsistent with the APA's own view of the distinction between standards and guidelines (see discussion above). The *Test Standards* are presented, and function, like APA's definition of guidelines. Future editions may want to reconcile this disparity.

**Table 2.1** Sources of validity evidence presented in standards and guidelines

<b>AERA/NCME/APA test standards</b>
Test content
Response processes
Internal structure
Relations to other variables
Consequences
<b>FDA</b>
Content validity
Other validity:
(a) Construct, (b) Convergent, (c) Discriminant, (d) Known-group, and (e) Criterion
<b>COSMIN</b>
Content validity
Structural validity
Cross-cultural validity
Criterion validity
<b>EMRPO</b>
Content-related
Construct-related
Criterion-related
<b>SIOP</b>
Evidence based on the relationship between scores on predictors and other variables
Content-related evidence
Evidence based on the internal structure of the test
Evidence based on response processes
Evidence based on consequences of personnel decisions
<b>Mental measurement yearbook</b>
Follows the AERA/APA/NCME <i>Standards for Educational and Psychological Testing</i>
<b>EFPA</b>
Construct validity
Criterion validity:
(a) Post-dictive or retrospective validity; (b) Concurrent validity; (c) Predictive validity

**FDA Guidance for Industry**

The Food and Drug Administration (FDA) of the United States published a document “*Guidance for Industry - Patient-Reported Outcomes Measures: Use in Medical Product Development to Support Labeling Claims*” (2009) on its current thinking regarding the review and evaluation of newly developed, modified, or existing patient-reported outcome (PRO) instruments for supporting labeling claims. Labeling claims are medical product labels constituting the formal approval of the benefits and risks of medical products by the FDA. The FDA defines PRO as “any report of the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else” (p. 2) and PRO instruments are means to “capture PRO data used to measure *treatment benefits* [italics in original] or risk in medical product clinical trials”

(p. 1). There is empirical evidence showing that a lack of validity evidence is one reason for PRO labeling claim rejection by the FDA (DeMuro et al. 2012). Therefore, ensuring that PRO instruments possess strong validity evidence is not inconsequential.

In reviewing and evaluating the quality of PRO instruments for labeling, the FDA takes into consideration a number of issues, including the usefulness of the PRO for the target patient population and medical condition, the design and objectives of the clinical studies, data analysis plans, the conceptual framework of the PRO instruments, and the measurement properties of the PRO instruments. The sources of validity evidence recommended by the FDA include content, construct, convergent, discriminant, known-group, and criterion. In the document, content validity is defined as the extent to which the PRO instrument measures the concept of interest. Evidence to support content validity of PRO instrument scores include item generation procedures, data collection method, mode of administration, recall period, response options, format and instructions, training related to instrument administration, patient understanding, scoring procedures, and respondent and administrator burden. Content validity evidence needs to be established before other measurement properties are examined and other properties such as construct validity or reliability cannot be used in lieu of content validity.

The FDA also recommends the inclusion of construct, convergent, discriminant, known-group, and criterion validity evidence to support the use of PRO for labeling claims. Construct validity is defined in the document as the extent to which the relations among items, domains, and concepts support a priori hypotheses about the logical relations that should exist with other measures. Convergent, discriminant, and known-group (the ability of a PRO instrument to differentiate between patient groups) validity are the sources of evidence to support construct validity. If appropriate, criterion validity, defined as the extent to which the scores of a PRO instrument correlate well with a “gold standard”, should also be examined. However, as PRO is used when one is measuring a concept that is best known from the patient perspective, therefore criterion validity evidence for most PRO instruments “is not possible because the nature of the concept to be measured does not allow for a criterion measure to exist.” (p. 20).

## **Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN)**

Developed by Mokkink and colleagues (2010b), the purpose of the *Consensus-based Standards for the selection of health Measurement Instruments* (COSMIN) checklist is to reach international consensus on the sources of measurement evidence that should be evaluated and to establish standards for evaluating the methodological quality (design requirements and preferred statistical procedures) of studies on measurement properties of psychometric instruments in health. The

checklist can also serve as a guide to the development and reporting of the measurement properties of health measurement instruments and academic journal editors and reviewers can use the checklist for appraising the methodological quality of measurement articles. It is important to note that the evaluation focus is on methodological quality, not on the quality of an instrument (Mokkink et al. 2010b). The checklist is primarily for PRO instruments but the checklist can also be used to evaluate the methodological quality of measurement properties studies of clinical rating and performance-based instruments. The taxonomy, terminology, and measurement properties definitions for the COSMIN checklist items have reached international consensus (Mokkink et al. 2010c). A manual is made publicly available to guide the use of the checklist.

The Delphi method (involving a group of experts participating in several rounds of surveys) was used to develop the COSMIN checklist. Four rounds of surveys were conducted between 2006 and 2007. International (majority of them from North America (25 people) and Europe (29 people) interdisciplinary experts (including psychologists, statisticians, epidemiologists, and clinicians) participated in the Delphi study. A total of 91 experts were invited and 57 (63 %) participated. Forty-three (75 %) of the 57 experts participated in at least one round of the Delphi and 20 (35 %) completed all four rounds. The experts had an average of 20 years (ranging from 6 to 40 years) of experience in health, educational, or psychological measurement research. Items on the final version of the COSMIN checklist are based on the consensus reached in the Delphi activities. The checklist contains ten categories, including (1) internal consistency, (2) reliability, (3) measurement error, (4) content validity (including face validity), (5) structural validity, (6) hypothesis testing, (7) cross-cultural validity, (8) criterion validity, (9) responsiveness, (10) interpretability. As presented in Table 2.1, the sources of validity evidence included in the COSMIN checklist include content validity and construct validity (which is subdivided into structural validity, hypothesis testing, and cross-cultural validity), and criterion validity.

A group of 88 raters from a number of countries (over half of them from the Netherlands) participated in the inter-rater agreement study for the COSMIN checklist. The mean number of years of experience in measurement research was nine, with a standard deviation of 7.1. The COSMIN checklist was used to rate a randomly selected 75 articles from the Patient-Reported Outcome Measurement (PROM) Group database, located in Oxford, United Kingdom. Each of the articles was rated by at least two raters (ranging from two to six raters). Inter-rater agreements for the COSMIN checklist items were satisfactory, with an agreement rate of over 80 % for two thirds of the checklist items (Mokkink et al. 2010a).

## Evaluating the Measurement of Patient-Reported Outcomes (EMPRO)

The *Evaluating the Measurement of Patient-Reported Outcomes* (EMPRO) tool is a 39-item instrument aimed at assessing the conceptual and theoretical models, psychometric properties, and administration procedures of PRO instruments and at assisting the selection of PRO instruments (Valderas et al. 2008). The development of the EMPRO began when the Spanish Cooperative Investigation Network for Health and Health Services Outcomes Research (Red IRYSS) was formed in 2002. One of the goals of the Red IRYSS was to promote the use of PRO instruments in the Spanish-speaking populations by developing an instrument for the standardized evaluation of characteristics of PRO instruments. The contents of the EMPRO items were based on the recommendations by the Medical Outcomes Trust (Scientific Advisory Committee of the Medical Outcomes Trust 2002).

In the development of the EMPRO, four experts were nominated and formed the panel (individuals with substantial knowledge and experience in the development, evaluation, and use of PRO). The panel experts generated the items for the EMPRO. The response formats and structure were based on the criteria for evaluating the quality of clinical guidelines by the AGREE Collaboration (2003). The final items were reviewed by a group of researchers on their contents, ease of use, and comprehensiveness.

The 39 items on the EMPRO covers eight categories, including (1) conceptual and measurement model, (2) reliability, (3) validity, (4) responsiveness, (5) interpretability, (6) burden, (7) alternative modes of administration, and (8) cultural and language adaptations and translations (see Table 2.1). EMPRO defines validity as the degree to which the PRO instrument measures what it claims to measure. The validity section of the EMPRO covers content (relevance, comprehensiveness, and clarity of items, and involvement of expert panels and target populations), criterion-related (association between the PRO instrument scores and a “gold standard” criterion), and construct evidence (hypotheses concerning the logical associations with other instruments and known-group differences). Table 2.1 presents the sources of validity evidence covered in EMPRO.

The EMPRO possesses satisfactory internal consistency, with a Cronbach’s alphas (for each of the eight categories) ranging from .71 to .83 and an overall alpha of .95. Inter-rater agreement rate was strong, with intra-class correlations (ICC) ranging between .87 and .94. A user’s manual and SPSS scoring algorithm for the EMPRO are available from Jose Valderas upon request.



## Principles for the Validation and Use of Personnel Selection Procedures

The *Principles for the Validation and Use of Personnel Selection Procedures* is the official guidelines by the Division 14 (Society for Industrial and Organizational Psychology [SIOP]) of the APA. Nancy Tippins, the then president of SIOP, formed a task force in 2000 to update the guidelines to make them consistent with the *Standards for Educational and Psychological Testing* (AERA et al. 1999) and with the current body of research. The purpose of the guidelines is to:

Specify established scientific findings and generally accepted professional practice in the field of personnel selection psychology in the choice, development, evaluation, and use of personnel selection procedures designed to measure constructs related to work behavior with a focus on the accuracy of the inferences that underlie employment decisions. (p. 1)

The guidelines are for procedures for personnel selection. Personnel selection procedures are defined as any procedure used to guide personnel selection decisions. These decisions often influence an individual's employment status and involve issues such as hiring, training, promotion, compensation, and termination. Personnel selection procedures include the use of, among others, traditional paper-and-pencil instruments, computer-based or computer-adaptive instruments, work samples, personality and intellectual assessment tools, projective techniques, individual biographical data, job interviews, reference checks, education and work experience, physical requirements, and physical ability assessment, singly or in combination.

As is the case in the *Standards for Educational and Psychological Testing* (AERA et al. 1999), the *Principles for the Validation and Use of Personnel Selection Procedures* recommends gathering and accumulating the same five sources of evidence to support the validity of score inferences for personnel selection decision making. The five evidence sources include (1) content-related evidence, (2) evidence based on the relationship between scores on predictors and other variables, (3) evidence based on the internal structure of the instrument, (4) evidence based on response processes, and (5) evidence based on consequences of personnel decisions. Table 2.1 presents the sources of validity evidence discussed in the document.

The first source of evidence, content-related evidence, concerns the degree of match between the content of a selection procedure and work content (which includes the work requirements or outcomes), as well as the format and wording of items or tasks, response formats, and guidelines regarding administration and scoring procedures. Evidence based on the relationship between scores on predictors and other variables can be obtained by demonstrating the association between two or more personnel selection procedures measuring the same (i.e., convergent validity) of distinct construct of interest (i.e., discriminant validity). Concurrent (predictor and criterion data collected at the same time) and predictive validity (the degree to which the scores of a selection procedure predict future job-related performance) evidence can also be gathered to support the evidence

based on relationship between scores on predictors and other variables. Evidence based on the internal structure of a personnel selection procedure involves the degree to which the items or tasks of a personnel selection procedure relate to each other, which supports the degree to which the items or tasks represent the construct of interest. Evidence based on response processes refers to the thinking processes involved when individuals give responses to items or tasks on selection procedures. This source of evidence can be gathered by asking respondents about their response strategies. Finally, evidence based on consequences of personnel decisions concerns the degree to which the intended use and misuse of selection procedures weakens the inferences. Group differences in the performance on selection procedures resulting in a disproportionate number of candidate being selected is an example of negative consequence (Zumbo 1999).

## **Buros Center for Testing: Mental Measurement Yearbook**

With a history of over 75 years, the Mental Measurement Yearbook (MMY) is an annual publication on reviews of measurement properties of commercially available tests in education and psychology. The idea began when Oscar Buros was receiving his graduate training at Columbia University, with an eye towards improving the quality of test manuals, as well as improving the science and practice of educational and psychological testing. The review process of the MMY is rigorous and the reviews provide test users with authoritative, accurate, and complete information regarding the quality of educational and psychological tests. The first MMY was published in 1938 and it is now published by the Buros Institute at the University of Nebraska, United States.

Each year the Buros Institute intends to include in the MMY commercially available tests in the English language that have not been previously reviewed and published in MMY. The Buros Institute maintains a working relationship with test publishers internationally and makes contacts to invite publishers to participate in the review process by submitting complementary test materials for review. Test publishers are not required to participate but doing so is a good professional practice (i.e., engaging external experts in various stages of test development) as stated in the *Standards for Educational and Psychological Testing* (AERA et al. 1999).

The MMY review model contains a number of sections, including (1) description of the test (e.g., intended purposes, target population, intended uses, administrative procedures), (2) development process (e.g., theoretical background, item development and selection, pilot testing), (3) technical details including standardization, reliability, and validity, (4) commentary (on the overall strengths and weaknesses of the test), (5) and summary (conclusions and recommendations) (see Table 2.1). The validity section of the Buros' MMY suggests information on:

Interpretations and potential uses of test results are addressed. Evidence bearing on valid uses of test scores may take the form of summarizing procedures or studies designed to investigate the adequacy of test content and testing measures. Evidence to support the use of test results to make classifications or predictions, where applicable, is described in this section. Differential validity of test score interpretation and use across gender, racial, ethnic, and culture groups should be examined. Comments concerning the quality of the evidence may be offered. (p. 130)

The review process at the Buros Institute follows most current edition of the *Standards for Educational and Psychological Testing* (AERA et al. 1999) and most of the tests are reviewed by two reviewers. The majority of the test reviewers reviewing tests and publishing reviews in the MMY possesses a doctoral degree and has taken courses in measurement. The Buros Institute has a database of over 900 test reviewers globally and test reviewers.

## EFPA Review Model

The European Federation of Psychologists' Association (EFPA) presented a model to systematically evaluate the quality of assessment instruments in education and psychology. The main objective is to provide test users with detailed, necessary information and rigorous evaluation about the quality of assessment instruments in education and psychology. The Task Force of the EFPA Broad consisting of 24 members was formed and the model was produced from a synthesis of a number of existing sources in Europe, including, among others, the Test Review Evaluation Form by the British Psychological Society and the Dutch Rating System for Test Quality. Table 2.1 presents the sources of validity evidence included in EFPA review.

In the EFPA model, it is stated that:

In the last decades of the past century, there was a growing consensus that validity should be considered as a *unitary concept* [emphasis added] and that differentiations in types of validity should be considered as different types of gathering evidence only. ... It is considered that construct validity is the more fundamental concept and that evidence on criterion validity may add to establishing the construct validity of a test. (p. 285)

Although the unitary view is mentioned, in the EFPA review model two sources of validity evidence are emphasized, including construct and criterion validity. It is stated that “the distinction between construct validity and criterion validity as separate criteria is maintained ... Construct-related evidence should support the claim that the test measures the intended trait or ability” (p. 288). A wide variety of research designs and statistical approaches can be used to gather construct validity evidence, including factor analysis (both exploratory and confirmatory), item-test correlations, measurement invariance, differential item functioning (DIF), multitrait-multimethod design, item response theory (IRT), experimental and quasi-experimental designs.

With respect to criterion validity, evidence is needed to demonstrate that “a test score is a good predictor of non-test behavior or outcome criteria” (p. 289). Criterion validity in the EFPA model includes (a) post-dictive or retrospective validity (focusing on the past), (b) concurrent validity (“same moment in time”), and (c) predictive validity (focusing on the future). The quality of the criterion “is dependent both on the reliability of the measure and the extent to which the measure represents the criterion construct” (p. 289). Although the EFPA model suggests that all tests require criterion validity evidence showing the strength of the relationships between a test and its criterion, strategies such as correlation-based analyses and sensitivity and specificity analyses can be used to establish criterion validity evidence. However, criterion validity may not be applicable if a test is not designed for prediction purposes (for example, a test aimed at measuring progress).

## **Do the Standards and Guidelines Reflect Contemporary Views?**

The extent to which the sources of validity evidence discussed in the seven standards and guidelines standards, guidelines are in line with the contemporary views of validity was examined. Content validity and association with other variables are discussed in all seven standards and guidelines. Internal structure is also discussed in the majority of the documents. Response processes and consequences are discussed only in the *Test Standards*, SIOP, and MMY. The *Test Standards*, SIOP, and MMY are the ones promoting that the various sources of validity evidence accumulated and synthesized are to support the construct validity of the scores of an instrument. This is not surprising given that SIOP and MMY following the APA, AERA, and NCME’s (1999) *Standards for Educational and Psychological Testing* and the *Test Standards* are heavily influenced by, among others, the work of Messick (1989).

## **Discussion**

In this chapter, an overview of the standards and guidelines relevant to the validation of measurement instruments for use in a number of disciplines (including business, education, health, and psychology) is provided. The extent to which these standards and guidelines reflect contemporary views of validity is also examined.

In contemporary views of validity, construct validity is the focus of validity. Various sources of evidence are accumulated and synthesized to support the construct validity of the interpretation and use of instruments. Close to half of the standards and guidelines reviewed in this chapter refer to the various sources of validity evidence as distinct types. These standards and guidelines suggest

validation practices as “stamp collecting” activities in which the different sources (and possibly one a single source) of validity can be collected to support the inferences drawn from instrument scores, without emphasizing the importance of the synthesis of various sources of evidence to support the construct validity of score inferences.

Response processes and consequences are only discussed in less than half of the standards and guidelines included in this review. Response processes are the investigation of the cognitive/thinking processes involved when an individual give responses to items. The purpose is not to examine people’s understanding of items, but rather to examine *how* and *why* people respond to items the way they do. Consequences refer to the intended use and misuse of instruments and are emerging as one of the main sources of validity evidence in today’s validation work (Hubley and Zumbo 2011, 2013). An example of consequences in validation is the use of screening tools for the diagnosis of clinical depression. The intended use of a screening tool is not to make official diagnosis, but to help clinicians identify individuals who may suffer from clinical depression and to identify those who may benefit from additional assessment to confirm an official diagnosis of clinical depression. Using the scores from a screening tool to make a diagnosis is an example of misuse. Misuse may have negative consequences on issues such as diagnostic decisions, insurance coverage, and epidemiology findings. It is however, important to note that the misuse of the instrument in and of itself does not invalidate the appropriate use of the instrument (in this case, for screening purposes). Rather, it is the use of the screen instrument for diagnostic purposes that make the score inferences invalid.

The quality of the validity evidence is also important. Some standards and guidelines included in this chapter suggest that we should not just focus on evaluating whether validity evidence exists, we should also pay attention to the methodological approaches employed to obtain the evidence. For instance, the EFPA model provides suggestions on the use of advanced statistical approaches such as item response modeling, measurement invariance analysis, and differential item functioning analysis to support internal structure. The COSMIN checklist is another good example of the evaluation of the methodological quality of studies conducted to support the validity of instrument scores.

The fact that contemporary views of validity have not penetrated all disciplines may be a reflection of a lack of impact of the modern views of validity on some disciplines such as health. It is also possible that the “one-shoe-fits-all” approach to validation may not work in the validation work across all disciplines. Standards and guidelines that are suitable for one discipline may not be applicable in the other. For instance, consequences may be particularly important in diagnostic tests and high-stakes educational assessment, whereas accumulating content validity evidence may be more important for obtaining FDA approval.

Individuals conducting validation work are encouraged to develop and situate a validation plan within the view of validity that is most suitable for the inferences one intends to make. A validation plan and the subsequent validity evidence accumulated and synthesized provide reviewers and authorities to judge the

strengths and appropriateness of the methodological approaches employed to obtain the evidence. See Chap. 19 of this edited book on our recommendations for validation practices.

**Acknowledgement** I thank Professor Bruno Zumbo for comments and suggestions.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for education and psychological testing*. Washington, DC: American Psychological Association.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Psychological Association. (1952). Committee on test standards. Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist*, 7, 461–465.
- American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51, 201–238.
- American Psychological Association. (2002a). Criteria for practice guideline development and evaluation. *American Psychologist*, 57, 1048–1051.
- American Psychological Association. (2002b). Criteria for evaluating treatment guidelines. *American Psychologist*, 57, 1052–1059.
- 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.
- American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- Carlson, J. F., & Geisinger, K. F. (2012). Test reviewing at the Buros Center for Testing. *International Journal of Testing*, 12, 122–135.
- Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale: Lawrence Erlbaum Associates.
- DeMuro, C., Clark, M., Mordin, M., Fehnel, S., Copley-Merriman, C., & Gnanasakthy, A. (2012). Reasons for rejection of patient-reported outcome label claims: A compilation based on a review of patient-reported outcome use among new molecular entities and biologic license applications, 2006–2010. *Value in Health*, 15, 443–448.
- Eccles, M. P., Grimshaw, J. M., Shekelle, P., Schünemann, H. J., & Woolf, S. (2012). Developing clinical practice guidelines: Target audiences, identifying topics for guidelines, guideline group composition and functioning and conflicts of interest. *Implementation Science*, 7, 60.
- European Medicines Agency, Committee for Medicinal Products for Human Use. (2005). *Reflection paper on the regulatory guidance for the use of Health-Related Quality of Life [HRQL] measures in the evaluation of medicinal products*. London: Author.
- Evers, A., Muñoz, J., Hagemester, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, 25, 283–291.

- Food and Drug Administration (2009) Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. Rockville: Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research.
- Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology*, 123, 207–215.
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103, 219–230.
- Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 1, pp. 3–19). Washington, DC: American Psychological Association Press.
- International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, 1, 93–114.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education/Praeger.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.
- Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.
- Mokkink, L. B., Terwee, C. B., Gibbons, E., Stratford, P. W., Alonso, J., Patrick, D. L., Knol, D. L., Bouter, L. M., & De Vet, H. C. W. (2010a). Inter-rater agreement and reliability of the COSMIN (Consensus-Based Standards for the Selection of Health Measurement Instruments) checklist. *BMC Medical Research Methodology*, 10, 82.
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & De Vet, H. C. W. (2010b). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, 19, 539–549.
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010c). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63, 737–745.
- Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of Life Research*, 11, 193–205.
- Shekelle, P. G., Woolf, S. H., Eccles, M., & Grimshaw, J. (1999). Clinical guidelines: Developing guidelines. *British Medical Journal*, 318, 593–596.
- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green: Author.
- The AGREE Collaboration. (2003). Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: The AGREE project. *Quality and Safety in Health Care*, 12, 18–23.
- Valderas, J. M., Ferrer, J., Mendivil, M., et al. (2008). Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. *Value in Health*, 11, 700–708.
- Woolf, S. H., Grol, R., Hutchinson, A., Eccles, M., & Grimshaw, J. (1999). Clinical guidelines: Potential benefits, limitations, and harms of clinical guidelines. *British Medical Journal*, 318, 527–530.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.

- Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Handbook of statistics, Vol. 26, pp. 45–79). Amsterdam: Elsevier.
- Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Charlotte: IAP – Information Age Publishing.



Validity and Validation in Social, Behavioral, and Health  
Sciences

Zumbo, B.D.; Chan, E.K.H. (Eds.)

2014, XV, 327 p. 15 illus., 2 illus. in color., Hardcover

ISBN: 978-3-319-07793-2