Social Indicators Research Series 54

Bruno D. Zumbo
Eric K.H. Chan *Editors*

Validity and Validation in Social, Behavioral, and Health Sciences

Springer

Validity and Validation in Social, Behavioral, and Health Sciences

Social Indicators Research Series

Volume 54

This new series aims to provide a public forum for single treatises and collections of papers on social indicators research that are too long to be published in our journal *Social Indicators Research*. Like the journal, the book series deals with statistical assessments of the quality of life from a broad perspective, It welcomes the research on a wide variety of substantive areas, including health, crime, housing, education, family life, leisure activities, transportation, mobility, economics, work, religion and environmental issues. These areas of research will focus on the impact of key issues such as health on the overall quality of life and vice versa. An international review board, consisting of Ruut Veenhoven, Joachim Vogel, Ed Diener, Torbjorn Moum, Mirjam A.G. Sprangers and Wolfgang Glatzer, will ensure the high quality of the series as a whole.

For further volumes:
http://www.springer.com/series/6548

Bruno D. Zumbo • Eric K.H. Chan
Editors

Validity and Validation in Social, Behavioral, and Health Sciences

Springer

Editors
Bruno D. Zumbo
Measurement, Evaluation, and Research Methodology (MERM) Program
Department of Educational and Counseling Psychology, and Special Education
The University of British Columbia
Vancouver, BC, Canada

Eric K.H. Chan
Measurement, Evaluation, and Research Methodology (MERM) Program
Department of Educational and Counseling Psychology, and Special Education
The University of British Columbia
Vancouver, BC, Canada

ISSN 1387-6570 ISSN 2215-0099 (electronic)
ISBN 978-3-319-07793-2 ISBN 978-3-319-07794-9 (eBook)
DOI 10.1007/978-3-319-07794-9
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014946725

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In this edited volume, 15 research syntheses of the validity evidence reported in different research areas are presented. The chapters were purposefully chosen to reflect a wide variety of disciplines, journals, or measures. Eight of the chapters focused on particular journals ranging from measurement and assessment journals like *Educational and Psychological Measurement*, *Psychological Assessment*, to international counterparts such as the *European Journal of Psychological Assessment*, as well as *Social Indicators Research: An International and Interdisciplinary Journal for Quality of Life Measurement*. In total 11 journals in a variety of disciplines that were North American, European, or International focused were surveyed in the chapters. From these journals one can see the far reach that we aimed to contain. Likewise, nine chapters focused on key tests, measures, or assessment tools that provide a sense of validation practices in particular areas of assessment. Note that one chapter focused both on a group of journals as well as particular measures. In short, in essence, we are studying the scholarly genre of validation reports and how this genre frames validity theory and practices.

Each chapter is meant to stand alone and hence one could read a sub-set of the chapters in any order. The "free-standing" nature of the chapters is important because readers may want to focus on one, or more chapters, because of the vast array of domains, topics, and measures we covered.

We were mindful that we wanted each chapter to be both unique but also use some common framework. Therefore, we decided that all chapters would, at least, follow the generic framework in the *Standards* (AERA et al. 1999) wherein five sources of validity evidence were of focus: (a) content-related, (b) response processes, (c) internal structure, (d) associations with other variables, and (e) consequences. The syntheses also addressed whether recent work in validity theory was cited as informing the validation practice (e.g., Hubley and Zumbo 1996, 2011, 2013; Kane 2006; Messick 1989; Zumbo 2007, 2009).

This volume represents a broad sampling of educational, psychosocial, and health research settings, giving us an extensive evidential basis to build upon earlier studies by Cizek and his colleagues (2008, 2010). It is worth noting that the chapters in this volume commonly used a sampling of papers because unlike Cizek et al. (2010) who

used a word search and hence were able to include hundreds of papers, the authors herein based their synthesis on a close read of the papers and not an automated word search. Therefore, in our authors' cases, the number of papers is limited by the methodology. This methodology has the benefit of contextualizing the findings reported in each of the papers being synthesized, and overall there are hundreds of papers (more than 500) reviewed in detail.

We would like to outline for you the general principles and ethos of the book. The book is organized in five parts. Part I consists of an introductory chapter that sets the stage for and purposes of the book, and a second chapter reviewing standards and guidelines for validation practices in a variety of academic disciplines and jurisdictions. Part II includes three chapters devoted to quality of life, wellbeing, and life satisfaction. Part III consists of six chapters broadly reflecting psychology and education. Part IV consists of six chapters in the broad domains of health and medicine, including health psychology, patient-reported outcomes, and medical education. It should be noted that the chapters in Parts II–IV overlap a great deal in focus (which is not surprising given the overall purpose of the book) and could be re-arranged with different section headings. The closing part includes two concluding chapters. The first is a "meta-synthesis" of the 15 research syntheses and the closing chapter takes the reader back to the broad focus of the whole volume.

Because of its breadth of scope and purpose, this book is a high watermark in the history of measurement, testing, and assessment because it documents what people do when they validate their tests, measures, or assessment instruments in a wide variety of disciplines and regions of the world. This focus on validation practices is interesting in and of itself and will influence both future validation studies and theorizing in validity. In part, it documents how validity theory is influencing validation practices, and it also guides us in developing a plan for validation work. In broad terms, we aimed to answer the question: What passes as validity evidence? In other words, when people validate a measure, what do they do? What does the academic community accept as evidence of measurement validity in its scholarly journals? It is important to note that our focus was not on whether the score inferences drawn from any particular measure, test, or assessment are "valid" but rather on the sources and kinds of validity evidence that are reported in the published research literature.

Like all studies, there are limitations to our work; the largest one is by design. Our focus is on papers published in scholarly journals. We did not include any synthesis of what testing organizations, testing companies, or professional test publishers are doing in their validation practices as reflected in test manuals or validation studies within their organizations. Some of this is captured in the work of Cizek and his colleagues (2008) in their study of the *Mental Measurement Yearbook*[1]; however, some of this information is also difficult to obtain because several testing organizations treat their validation studies as propriety information. As a reminder, however, our focus was on papers published in scholarly journals, and as

[1] Curiously, their overall findings are consistent with ours.

we show in our search of the PsycInfo database in Figs. 1.1, 1.2, and 1.3, we have a large body of work to select from and hence our focus is warranted.

We would like to close by acknowledging the impressive body of work that our collaborators amassed. To support the reading of each chapter, each chapter author was asked to speak to validity theory in their domain and, where possible, make recommendations for validation practices. There is much gold to be mined for validity theorists and practitioners in the closing sections of each chapter. In addition to our own review of each of the chapters, we would like to thank Dr. Katie Gunnell, Dr. Rebecca (Beck) Collie, Michelle (Yue) Chen, and Dr. Dallie Sandilands who each provided valuable feedback for several chapters.

Vancouver, BC, Canada

Bruno D. Zumbo
Eric K.H. Chan

References

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.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*, 397–412.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement, 70*, 732–743.

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.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education/Praeger.

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

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.

Contents

Contributors

Neelam Ark Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Tavinder K. Ark Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Eric K.H. Chan Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Michelle Y. Chen Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Mary L. Chinni Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Rebecca J. Collie School of Education, University of New South Wales, Sydney, NSW, Australia

Daniel W. Cox Counseling Psychology Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Peter R.E. Crocker School of Kinesiology, The University of British Columbia, Vancouver, BC, Canada

Ira Darmawanti Department of Educational Psychology and Guidance, State University of Surabaya, Surabaya, East Java, Indonesia

Anne M. Gadermann Centre for Health Evaluation and Outcome Sciences, St. Paul's Hospital, and The University of British Columbi, Vancouver, BC, Canada

Katie E. Gunnell School of Kinesiology, The University of British Columbia, Vancouver, BC, Canada

Alexander H.S. Huang Counseling Psychology Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Anita M. Hubley Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Yan Liu Harvard Medical School, Harvard University, Boston, MA, USA

Juliette Lyons-Thomas Regents Research Fund, Institute for Urban and Minority Education, Teacher's College, Columbia University, New York, NY, USA

Diane E. Mack Behavioural Health Sciences Research Lab, Department of Kinesiology, Brock University, St. Catharines, ON, Canada

Hillary L. McBride Trinity Western University, Langley, BC, Canada

Marvin J. McDonald Trinity Western University, Langley, BC, Canada

Olievia P. Mulyana Department of Educational Psychology and Guidance, State University of Surabaya, Surabaya, East Java, Indonesia

David W. Munro Counseling Psychology Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Jess J. Owen Counseling Psychology Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Debra (Dallie) Sandilands Faculty of Education, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Ayumi Sasaki Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Benjamin J.I. Schellenberg Department of Psychology, The University of Manitoba, Winnipeg, MB, Canada

Benjamin R. Shear Graduate School of Education, Stanford University, Stanford, CA, USA

Roya Vojdanijahromi Counseling Psychology Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Rachel M. Wiens Trinity Western University, Langley, BC, Canada

Philip M. Wilson Behavioural Health Sciences Research Lab, Department of Kinesiology, Brock University, St. Catharines, ON, Canada

Wen Zhang Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Sophie Ma Zhu Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Bruno D. Zumbo Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, Vancouver, BC, Canada

Part I
Opening Section

Chapter 1
Setting the Stage for *Validity and Validation in Social, Behavioral, and Health Sciences*: Trends in Validation Practices

Bruno D. Zumbo and Eric K.H. Chan

As witnessed in the seminal work of Messick (1989) and Kane (2006, 2013), over the last 50 years validity theories have become more expansive and complex. Prior to the 1950s, a diversity of procedures was used in validation practice and an array of names for these procedures was used when researchers reported validity evidence. Early in the history of the social and behavioral sciences, the criterion- and content-based models dominated the practice of validation (Anastasi 1986). The early practices reflected the then dominant 'behavioral' view in the social sciences and hence tests and measures were primarily considered predictive devices – wherein one could predict some future behavior, or was a short-hand for a more complex current behavior. With this in mind, one can see how the correlation with the criterion (i.e., the future or current behavior) was the dominant perspective in validation. Simply put, a test or measure was valid if it predicted the criterion. In 1954, the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (the first version of the North American test standards) was published by the American Psychological Association in collaboration with the American Educational Research Association and the National Council on Measurement in Education. In this document, validity was classified into content, predictive, concurrent, and construct. A year later, Cronbach and Meehl (1955) published a seminal paper and argued that the focus should be on construct validity, emphasizing the importance of a nomological network as a form of theory building about the psychological phenomenon of interest. This signaled the change in viewing tests and measures as reflective devices (or signs) of some unobserved phenomena (i.e., one definition of a construct). This shift in emphasis to unobserved phenomena is an important landmark in the history of measurement,

B.D. Zumbo, Ph.D. (✉) • 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: bruno.zumbo@ubc.ca

B.D. Zumbo and E.K.H. Chan (eds.), *Validity and Validation in Social, Behavioral, and Health Sciences*, Social Indicators Research Series 54, DOI 10.1007/978-3-319-07794-9_1,

assessment, and testing. Please note, however, that the criterion view still continued but had less emphasis as the discipline of psychological theorizing began to dwell again among unobservables in response to the various forms of behaviorism that shun these unobservables.

Over three decades after Cronbach and Meehl (1955), Messick (1989) published a seminal paper on the unitary view of validity. According to Messick (1989), validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13) and is a fundamental concern in measurement. Messick's (1989) unitary view of validity remains influential in the theoretical arena of measurement and is reflected in the *Standards for Educational and Psychological Testing* (AERA et al. 1999). According to the *Standards*, validity is "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (p. 9). This perspective has given rise to the situation wherein there is no singular source of evidence sufficient to support a validity claim.

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). Validity is not about the instrument, test, or measure but rather about the inferences, claims, or decisions that one makes based on the scores. Therefore, one does not validate a test, measure, or assessment but rather one validates the inferences. Validity does not exist as distinct types and validation should not be a piecemeal activity akin to stamp collecting – or, for that matter, collecting baseball, soccer, or hockey cards. Validation is an ongoing process in which various sources of validity evidence are accumulated and synthesized to support the construct validity of the interpretation and use of instruments. In addition to the traditional sources of evidence such as content, relations to other variables (e.g., convergent, discriminant, concurrent, and predictive validity), and internal structure (dimensionality), evidence based on consequences (intended use, and misuse), and response processes (cognitive processes during item responding or during rating) are important sources of validity evidence that should be included in validation practices. Although different validity theorists emphasize each of these to varying amounts, validation practices center around establishing a validity argument (such as Cronbach and Kane), an explanation for score variation (such as Zumbo), a theoretical framework of law-like relations that is tested against data (a nomological network, Cronbach and Meehl), 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). As a whole, these foci capture the core perspectives on validity seen in the current literature and are meant to guide the practice of validation. It should be noted that, as expected in a vibrant scholarly discipline, elements of this contemporary view are not endorsed by all and, in fact, are challenged by some important voices in the field (e.g., Borsboom et al. 2004; Markus and Borsboom 2013).

Trends in Validation Practices: Setting the Stage

We conducted a systematic search of validation studies published since the 1960s. Our aim was to get a snapshot of the trends in validation practices for publications that explicitly presented themselves as validation studies. Of course, a good deal of validation work is done alongside substantive studies (wherein the substantive studies are the primary objective) in psychology, education, health, and other social and behavioral sciences, however, we wished to trace the validation practices of studies for which the validation work is the primary (if not sole) purpose of the publication. We did this because we believe that focusing on studies that are explicitly cast as validation studies will give us the clearest picture of validation practices. When one is doing validation as a side project to a larger study that one considers more substantive then the validation practices will likely be described in less detail and likely also a modest or minor part of the body of work. For example, if one is interested in the mediating and moderating factors in the relation between academic self-concept and academic achievement, one may report a small-scale validation exercise along the way but certainly, by definition, that validation study will be relatively limited in scope and the details presented in the manuscript as compared to a study that has as its sole purpose the reporting of a validation study.

We were interested in documenting the general trend in publication of validation studies. For each 5-year period between 1961 and 2010 we searched the PsycInfo database for the terms 'validity' or 'validation' and the terms 'psychometric', 'measurement', 'assessment' or 'test' in the abstract of the paper. In addition, we limited our search to peer-reviewed scholarly journals. As presented in Fig. 1.1 there is clearly an increase in the number of scholarly peer-reviewed journal publications with just over 300 publications between 1961 and 1965 to over 10,200 publications between 2006 and 2010. Certainly, some of that increase can be attributed to the increase in the sheer number of journals and researchers; however, the fact is that the field of measurement validity is growing in remarkable strides.

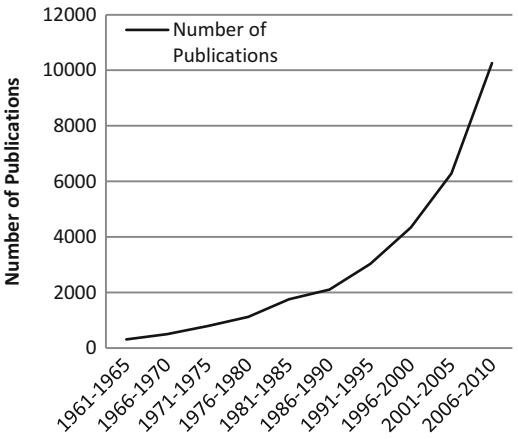

Fig. 1.1 Trend line depicting the pattern of publication of validation studies

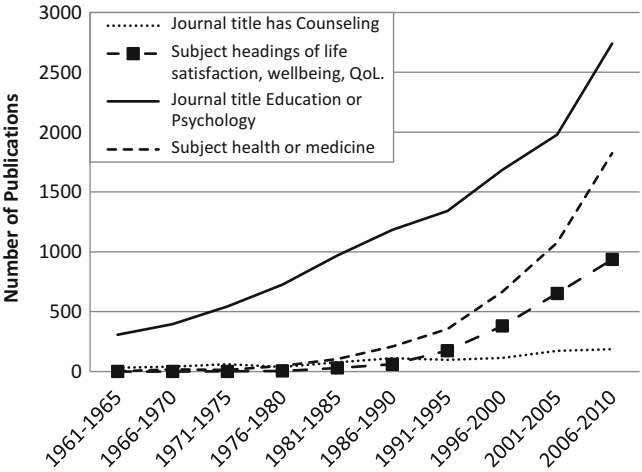

Fig. 1.2 Trend lines of publication of validation studies across disciplines

In Fig. 1.2 we documented the publication practices in four domains. Two of the trend lines represent well-established areas of measurement research that have journals dedicated to them: education or psychology, and counseling. The remaining two trend lines represent relatively emerging fields of measurement, testing, or assessment defined by terms such as 'life satisfaction, wellbeing, or quality of life (QoL)', and 'health or medicine'. Again, like Fig. 1.1, we are witnessing an increase in the number of scholarly publications in these disciplines with, as expected, the greatest increase being seen in education and psychology.

Once again, in Fig. 1.3 we applied the same search strategy except that in this case we searched for various sources of validity evidence. For example, in documenting the trend in content validation studies, we searched for the terms "content validity" or "content validation" and the terms 'psychometric', 'measurement', 'assessment' or 'test' in the abstract of the papers. We continued to limit our search to peer-reviewed scholarly journals. Noting, of course, that papers can report more than one source of validity evidence, construct validity evidence is the most commonly reported followed by concurrent and predictive evidence, and finally content validity evidence.

It is important to note that in the data reported in Figs. 1.1, 1.2 and 1.3 we are looking back in time with the labels from the current *Standards*. In essence, we are looking back over our shoulders but applying today's labels. Likewise, it is important to note that this is a "snapshot" picture that is obtained by documenting the count of words in the abstracts of the published articles and hence does not document the specifics, nor does it break it down by scholarly practices. In fact, it is this general picture that motivates the need for the studies reported in this edited volume.

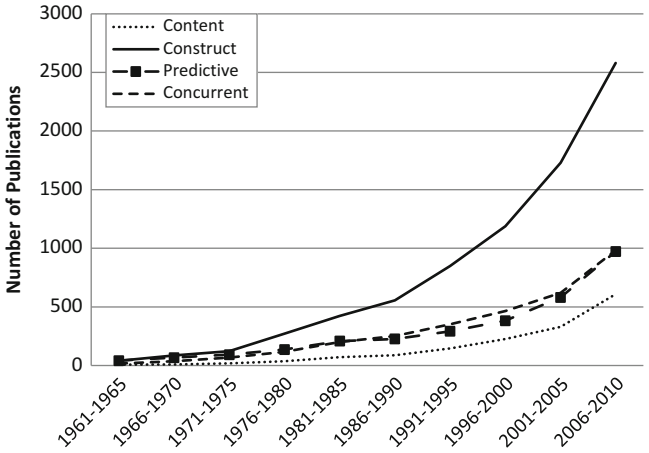

Fig. 1.3 Trend lines of publication of validation studies across sources of validity evidence

With the growing number of validation papers published in academic journals across different academic disciplines, and with the revision of the *Test Standards* scheduled to be released soon, it is timely to examine validation practices by researchers across different academic disciplines. Our focus, and the focus of this edited volume, is a study of the scholarly genre of validation reports and how this genre frames validity theory and practices.

References

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. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, 201–238.

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

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale: Lawrence Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. doi:10.1037/h0040957.

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.

Kane, M. T. (2006). Educational measurement. In R. L. Brennan (Ed.), *Validation* (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 and Macmillan.

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 Science B.V.

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.

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

B.D. Zumbo and E.K.H. Chan (eds.), *Validity and Validation in Social, Behavioral, and Health Sciences*, Social Indicators Research Series 54, DOI 10.1007/978-3-319-07794-9_2,

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)[1]
2. *Guidance for Industry – Patient-Reported Outcomes Measures: Use in Medical Product Development to Support Labeling Claims* (Food and Drug Administration 2009)[2]
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)

[1] 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.

[2] The European Medicines Agency (EMA 2005) published a document providing broad recommendations on the use of health-related qualify 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

AERA/NCME/APA test standards
Test content
Response processes
Internal structure
Relations to other variables
Consequences
FDA
Content validity
Other validity:
(a) Construct, (b) Convergent, (c) Discriminant, (d) Known-group, and (e) Criterion
COSMIN
Content validity
Structural validity
Cross-cultural validity
Criterion validity
EMRPO
Content-related
Construct-related
Criterion-related
SIOP
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
Mental measurement yearbook
Follows the AERA/APA/NCME *Standards for Educational and Psychological Testing*
EFPA
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).