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.

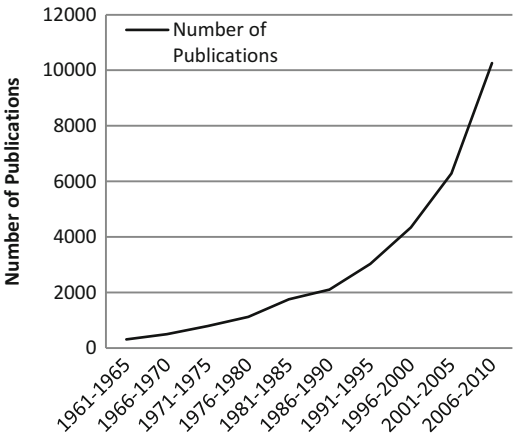

Fig. 1.1 Trend line depicting the pattern of publication of validation studies

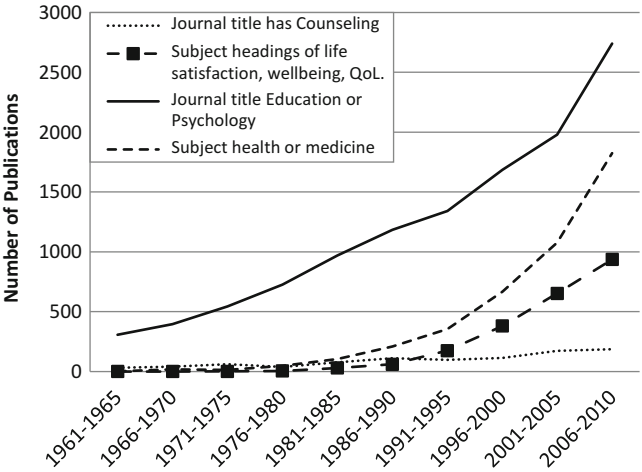

Fig. 1.2 Trend lines of publication of validation studies across disciplines

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.

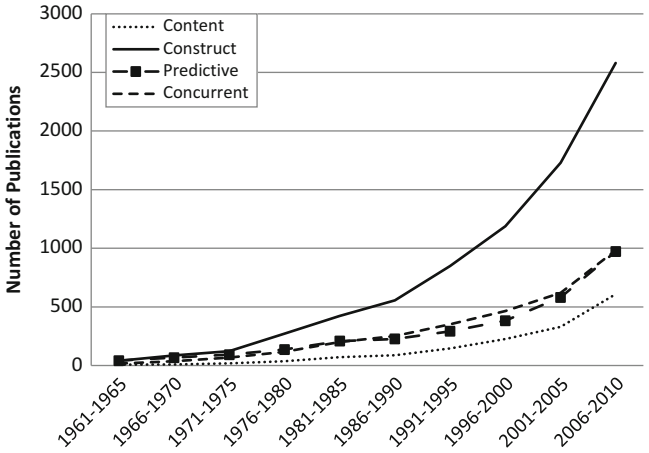

Fig. 1.3 Trend lines of publication of validation studies across sources of validity evidence

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).

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. Personal 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. Personal 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., Muñiz, J., Hagemeister, 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., Mendívil, 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.

Part II
Quality of Life, Wellbeing, and Life Satisfaction

Chapter 3
Reporting of Measurement Validity in Articles Published in *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*

Bruno D. Zumbo, Eric K.H. Chan, Michelle Y. Chen, Wen Zhang, Ira Darmawanti, and Olievia P. Mulyana

Quality of life (QoL) and social indicators research have become an area of major focus in the social, behavioral, and health sciences. The disciplines of QoL and social indicators research are truly trans-disciplinary and span psychology, sociology, health, education, economics, political science, and public policy. One can imagine that this broad span of disciplines would result in a variety of empirical approaches to measurement. Interestingly, two classes of measurement have evolved: psychometric and the economic utility traditions. The focus of the present chapter is to describe the validation practices in the psychometric tradition in the journal *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement* (henceforth referred to as *Social Indicators Research*). This journal is an important focus of study because of its history as the first journal focused on social indicators and QoL (founded in 1974) and because it has become the leading journal for the publication of research results dealing with measurement of the quality of life. The journal is interdisciplinary including papers on psychological well-being, health, education, the natural environment, social customs and morality, mental health, law enforcement, politics, economics, religion, and science and technology.

With such a wide focus an interesting question arises as to the validation practices reported in the journal. As Zumbo (1998) states, the concept, method, and process of validation is central to QoL and social indicators research, for without validation, any inferences made from a measure are meaningless.

B.D. Zumbo, Ph.D. (✉) • E.K.H. Chan • M.Y. Chen • W. Zhang
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

I. Darmawanti • O.P. Mulyana
Department of Educational Psychology and Guidance, State University of Surabaya, Ketintang Baru XIV/2, Surabaya, East Java 60231, Indonesia

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_3,

The journal has contributed both to the theoretical developments in validity theory and the practice of validation, as well as the dissemination of validation studies. Validation studies regularly appear in the journal and two special issues have been devoted to validity theory and validation practices. In 1998 Bruno D. Zumbo edited a combined three issue, over 500 page volume, Volume 45 Numbers 1–3, entitled "Validity Theory and the Methods Used in Validation: Perspectives from Social and Behavioral Sciences". That compendium included Samuel Messick's last paper before he passed away, and the entire volume is dedicated to him and his achievements. The second special issue appeared in 2011, Volume 103 Issue 2, approximately 146 pages, was edited by Martin Guhn, Bruno D. Zumbo, Magdalena Janus, and Clyde Hertzman and was entitled "Validation Theory and Research for a Population-Level Measure of Children's Development, Wellbeing, and School Readiness".

Through its regular publication of validation studies, and its special issues, *Social Indicators Research* has contributed to validity theory and validation methods becoming more complex and expansive over the past several decades. There is an agreement among validity experts that the accumulation and integration of evidence from different sources is needed to support the validity of the interpretation and inferences made from the scores arising from measurement instruments (AERA et al. 1999; Kane 2006; Messick 1989; Zumbo 2007, 2009). The contemporary view of validity contends that in addition to the traditional sources of validity such as content, relations to other variables (e.g., discriminant, and convergent validity), and internal structure, 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 validating psychometric instruments (AERA et al. 1999; Messick Hubley and Zumbo 2011, 2013).

The purpose of this study was to review the reporting of validity evidence and validation practices in papers published in *Social Indicators Research*. This effort serves three purposes: (1) To review validation practices in the area of QoL, (2) To investigate the gap between the theories of validity and the practices of validation in the area of QoL, and (3) to make recommendation for practice. It is important to keep in mind that the focus is on validation practices and not on whether the uses of or inferences from any particular instrument, measure, or assessment is valid, per se.

Method

We conducted a systematic search using the official website of the journal on February 23, 2013. We used "validity" or "validation" as the search keywords (resulted in over 1,000 hits, but not all of them were validity papers). Although it would be ideal to collect and review all validity and validation articles published in the journal, due to the large number of articles and the limited resources we have, we decided to only include papers published between 2012 and the date we

conducted our search, with an explicit focus on papers with the term "valid", "validity", or "validation" in the title. It should be noted that both published papers and those in the print queue (i.e., the online first papers) were included in our search. We believe our approach is a good way to capture papers that are explicitly stated as validity papers and allows us to document the recent practice of validation in this journal. A total of 24 articles were coded using our coding sheet. We included only empirical validation studies and excluded opinion and editorial articles, reviews, systematic reviews, and meta-analyses, theoretical papers, and articles on guidelines, and expository papers on statistical applications.

To code the characteristics and validity evidence presented in each study, a coding sheet was developed. Building from earlier research (Cizek et al. 2008, 2010) and using the validity evidence framework outlined in the most current version of the *Standards for Educational and Psychological Testing* (AERA et al. 1999), the sources of validity evidence included in our coding were: face, content, construct, predictive, concurrent, convergent, discriminant, response processes, consequences, reliability, and other. The sources of validity were coded based on what the authors reported and the procedures involved. In our coding, if for instance an author in the article explicitly reported "convergent validity" and presented empirical findings such as correlations between two instruments, convergent validity was coded. In instances where the author reported conducting "think-aloud" process to investigate people's responses yet did not explicitly call it "response processes", we still coded it as response process evidence. Similarly, if an author presented factor analytic results but did not call it internal structure, it was still coded as internal structure evidence.

The coding for this project was conducted by four of the authors of this paper. The coding for the 24 selected articles were distributed as follow: Each of the four individuals independently coded five articles (total = 20), and the remaining four articles were coded by all four individuals as agreement check. In other words, each of the four individuals coded nine (five unique and four common) articles. With respect to the agreement in the coding among the four raters, disagreement occurred in 15 of the 40 multiple ratings (i.e., 4 articles for the 10 sources of validity that were coded). Within 9 of the 15 cases where disagreement occurred, 3 of the raters agreed with each other and only 1 rater disagreed with the other 3 raters. Disagreements in the coding were reviewed and inconsistencies were resolved by the second author.

During coding, three studies (Bulloch 2013; Diener et al. 2013; Haeken and Munck 2012) were excluded because it was ascertained that they were review articles and not empirical reports of validation findings. As a result, the final total number of studies included in this study was 21.

Results and Conclusions

The papers included in our review typically did not frame their validation practices in terms of the AERA, APA, NCME *Test Standards* and only one cited the *Standards* and the work of contemporary validity theorists including Kane, Messick, and Zumbo.

The number of sources of validity evidence reported per study ranged from 1 to 6 (see Table 3.1). Table 3.2 presents the percentage of the sources of validity evidence reported in the articles we reviewed. Internal structure was the most frequently reported source of evidence, reported in three quarters of the papers. Half of the studies reported convergent validity evidence. Construct validity evidence (primarily examined using factor analysis) was reported in more than a third of the studies and face validity was reported in over a quarter of the studies. Examples of the methods employed to investigate face validity included face validity screening by core researchers, insights provided by school and community partners, and comparing test items with conceptual definition of the construct.

Slightly more than a fifth of the studies reported evidence on content validity and approximately fifteen percent of the papers reported predictive validity evidence. Methods employed to study content validity included seeking expert opinion, asking participants to provide feedback on the length of the instrument, the comprehensibility of the instructions, items, response options, and content, as well as

Table 3.1 Frequency of number of validity sources reported

Number of sources	Frequency	Percent
1	4	19.0
2	6	28.6
3	6	28.6
4	2	9.5
5	2	9.5
6	1	4.8
Total	**21**	**100**

Table 3.2 Sources of validity evidence reported[a]

Source of validity evidence	Number of papers	Percent of papers
Internal structure	16	76.19
Convergent	11	52.38
Construct	9	42.86
Face	6	28.57
Content	6	28.57
Discriminant	3	14.29
Predictive	3	14.29
Concurrent	2	9.52
Response processes	2	9.52
Consequences	0	0

[a]A paper can report more than one source of validity

scrutinizing items to make sure they reflect the construct of interest. Discriminant validity, which can serve as a baseline to compare convergent validity, was also reported in three of the papers. Concurrent validity and response processes were each reported in two of the studies. Examples of the methods employed to examine response processes included cognitive interviewing and face-to-face interviews using a "talk aloud" procedure to elicit the thoughts that went through people's mind when answering items. Consequences, a source of validity emerging as central to validity claims, were not reported in any study. Although the importance of response processes and consequences in validation have been well documented (Hubley and Zumbo 2011, 2013; Messick 1989, 1995; Zumbo 2007, 2009), these two sources are rarely presented in papers published in this journal.

The purpose of this study was to review the reporting of validity evidence in papers published in *Social Indicators Research*. For the most part, papers appeared to present validity evidence without an overarching plan for the validation or a justification of the validity evidence provided – 'validation as stamp collecting'. This is a concern because validation practices would be strengthened if the evidence was tied to the purpose of the measurement and was integrated in some way to achieve some sort of objective in the validation process. Likewise, the 1998 special issue of this same journal is not cited in the papers included in the present review. This lack of citation may be a reflection of sampling, a systematic bias in citations, or a lack of impact of that special issue. The latter is not supported given that the papers in the 1998 special issue have been cited 1,059 times, with an average of 50 citations per paper and 66 citations per year – as per Google Scholar. Overall, the conclusion can be drawn that the validation studies are reported without reliance on the theory of validity and validation, including the theory and methods introduced in this same journal.

We have three specific recommendations. First, it is important that one situates a validation plan within a view of validity (e.g., Messick 1989; Kane 2006, 2013) because doing so gives readers the vantage point of being able to judge the strength of the validity evidence, including both the purpose of the validation as well as a sense of what the current validation study has not presented. Second, we also urge more studies that focus on response processes. For example, Gadermann et al. (2011) used cognitive interviews and sought evidence of whether, for example, Michalos' multiple discrepancies theory (MDT; Michalos 1985) is evidenced in the response processes – wherein MDT can be considered an explanatory model of item responding (see Zumbo 2007, 2009). The recent paper by Castillo-Díaz and Padilla (2013) is exemplary because it demonstrates the power of cognitive interviewing in understanding the cognitive structure of item responding, and in addition it ties the evidence directly to contemporary validity theories by addressing the role of cognitive interviewing in Kane's argument-based approach (Kane 2013) and Zumbo's contextualized view of validity (Zumbo 2007, 2009). Investigating the process of item responding in such a manner closes the inferential gap and is strong evidence – what Messick refers to as the "substantive aspect" of construct validity evidence. Finally, researchers who are validating measures (or simply using measures or scales in their research) need to recognize

that values play a central role from framing the construct to choosing a psychometric or validation method. These values, and their implications for intended as well as unintended social and personal consequences of using a particular measure or scale, are fundamental to the measurement process and to validation (Hubley and Zumbo 2011).

References[1]

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.

*Anagnostopoulos, F., & Griva, F. (2012). Exploring time perspective in Greek young adults: Validation of the Zimbardo Time perspective inventory and relationships with mental health indicators. *Social Indicators Research, 106*, 41–59.

*Breheny, M., Stephens, C., Alpass, F., Stevenson, B., Carter, K., & Yeung, P. (2013). Development and validation of a measure of living standards for older people. *Social Indicators Research, 114*, 1035–1048.

Bulloch, S. L. (2013). Seeking construct validity in interpersonal trust research: A proposal on linking theory and survey measures. *Social Indicators Research, 113*, 1289–1310.

*Castillo-Díaz, M., & Padilla, J.-L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research, 114*, 963–975.

*Christensen, L. N., Ehlers, L., Larsen, F. B., & Jensen, M. B. (2013). Validation of the 12 item short form health survey in a sample from region central Jutland. *Social Indicators Research, 114*, 513–521.

*Christodoulou, C., Schneider, S., & Stone, A. A. (2014). Validation of a brief yesterday measure of hedonic well-being and daily activities: Comparison with the day reconstruction method. *Social Indicators Research, 115*, 907–917.

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.

*Dickes, P., & Valentova, M. (2013). Construction, validation and application of the measurement of social cohesion in 47 European countries and regions. *Social Indicators Research, 113*, 827–846.

Diener, E., Inglehart, R., & Tay, L. (2013). Theory and validity of life satisfaction scales. *Social Indicators Research, 112*, 497–527.

*Dinç, L., Korkmaz, F., & Karabulut, E. (2012). A validity and reliability study of the multidimensional trust in health-care systems scale in a Turkish patient population. *Social Indicators Research, 113*, 107–120.

*Fayad, Y. I., & Kazarian, S. S. (2013). Subjective vitality of Lebanese adults in Lebanon: Validation of the Arabic version of the subjective vitality scale. *Social Indicators Research, 114*, 465–478.

[1] References marked with an asterisk indicate studies included in this review.

*Fleury-Bahi, G., Marcouyeux, A., Préau, M., & Annabi-Attia, T. (2013). Development and validation of an environmental quality of life scale: Study of a French sample. *Social Indicators Research, 113*, 903–913.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction with Life Scale adapted for children: A focus on cognitive processes. *Social Indicators Research, 100*, 37–60.

Guhn, M., Zumbo, B. D., Janus, M., & Hertzman, C. (2011). Validation theory and research for a population-level measure of children's development, wellbeing, and school readiness. *Social Indicators Research, 103*, 183–191.

Haeken, A., & Munck, G. L. (2012). Cross-national indices with gender-differentiated data: What do they measure? How valid are they? *Social Indicators Research, 111*, 801–838.

*Henderson, L. W., Knight, T., & Richardson, B. (2014). The hedonic and eudaimonic validity of the orientations to happiness scale. *Social Indicators Research, 115*, 1087–1099.

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.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

*Lee, C.-J., & Kim, D. (2013). A comparative analysis of the validity of US state- and county-level social capital measures and their associations with population health. *Social Indicators Research, 111*, 307–326.

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

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

Michalos, A.C. (1985). Multiple discrepancies theory (MDT). *Social Indicators Research, 16*, 347–413.

*Oliva, A., Antolín, L., & López, A. M. (2012). Development and validation of a scale for the measurement of adolescents' developmental assets in the neighborhood. *Social Indicators Research, 106*, 563–576.

*Sancho, P., Galiana, L., Gutierrez, M., Francisco, E.-H., & Tomas, J. M. (2014). Validating the Portuguese version of the satisfaction with life scale in an elderly sample. *Social Indicators Research, 115*, 457–466.

*Schonert-Reichl, K. A., Guhn, M., Gadermann, A. M., Hymel, S., Sweiss, L., & Hertzman, C. (2013). Development and validation of the middle years development instrument (MDI): Assessing children's well-being and assets across multiple contexts. *Social Indicators Research, 114*, 345–369.

*Sen, S., Sen, G., & Tewary, B. K. (2012). Methodological validation of quality of life questionnaire for coal mining groups-Indian scenario. *Social Indicators Research, 105*, 367–386.

*Silva, A. J., & Caetano, A. (2013). Validation of the flourishing scale and scale of positive and negative experience in Portugal. *Social Indicators Research, 110*, 469–478.

*Tomyn, A. J., Norrish, J. M., & Cummins, R. A. (2013). The subjective wellbeing of indigenous Australian adolescents: Validating the personal wellbeing index-school children. *Social Indicators Research, 110*, 1013–1031.

*Vasconcelos-Raposo, J., Fernandes, H. M., Teixeira, C. M., & Bertelli, R. (2012). Factorial validity and invariance of the Rosenberg self-esteem scale among Portuguese youngsters. *Social Indicators Research, 105*, 483–498.

*Wang, N., Kosinski, M., Stillwell, D. J., & Rust, J. (2014). Can well-being be measured using Facebook status updates? Validation of Facebook's gross national happiness index. *Social Indicators Research, 115*, 483–491.

*Zhang, H., Xu, X., & Tsang, S. K. M. (2012). Conceptualizing and validating marital quality in Beijing: A pilot study. *Social Indicators Research, 113*, 197–212.

*Zhang, J., Lu, J., Zhao, S., Lamis, D. A., Li, N., Kong, Y., Jia, C., Zhou, L., & Ma, Z. (2014). Developing the Psychological Strain Scales (PSS): Reliability, validity, and preliminary hypothesis tests. *Social Indicators Research, 115*, 337–361.

Zumbo, B. D. (1998). Opening remarks to the special issue on Validity theory and the methods used in validation: Perspectives from the social and behavioral sciences. *Social Indicators Research, 45*, 1–3.

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.

Chapter 4
A Research Synthesis of Validation Practices Used to Evaluate the Satisfaction with Life Scale (SWLS)

Mary L. Chinni and Anita M. Hubley

Researchers are faced with a vast number of studies reporting a variety of results when exploring any topic. Olkin (1996) identified a roughly ten-fold increase in the number of research publications between 1940 and 1990 from 2,300 to 25,000 biomedical journals, 91 to 1,100 journals in psychology, and 91 to 920 journals in mathematics. The amount of information available creates a formidable challenge to researchers and practitioners needing to gather, assimilate, and critically assess the volume of scientific information available to them. Moreover, Cooper et al. (2009) suggested that the increasing volume of knowledge has led to a narrowing of specialties within scientific fields and thus an increasing reliance by researchers on literature reviews to stay current with developments in their fields.

The terms 'research synthesis', 'literature review', and 'systematic review' are often used interchangeably (Cooper 2010). A research synthesis can be thought of as a type of literature review whose primary intention is to assess the quality of information available, to determine whether research findings are consistent and generalizable across populations, and to determine the extent to which findings vary across studies and populations (Mulrow 1994). Manten (1973) adds that literature reviews are "not based primarily on new facts and findings, but on publications containing such primary information whereby the latter is digested, sifted, classified, simplified, and synthesized" (p. 75). What further distinguishes a research synthesis from a literature review is the specific identification of what is to be examined within a literature, and a methodology for examination that can be replicated. Key elements of a research synthesis include: (1) a clearly stated set of objectives, (2) pre-set eligibility criteria for articles used in the study, (3) a methodology that can be replicated, (4) a systematic search to identify studies that meet the eligibility criteria, (5) an assessment of the soundness of all findings, and

M.L. Chinni • A.M. Hubley (✉)
Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
e-mail: anita.hubley@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_4,

(6) a systematic presentation of the results of all studies included in the analysis (The Cochrane Collaboration 2002).

A research synthesis of validation practices seeks to examine the methods and procedures that researchers use to evaluate measures and determine whether inferences made about respondents based on those measures are appropriate. Validity is a fundamental concern to measurement specialists and practitioners who use tests to inform and justify social policy decisions, medical and psychological assessments, or an individual's placement, training, and licensing within educational and professional contexts. The *Standards for Educational and Psychological Testing*[1] (AERA et al. 1999) assert that validity is "the most fundamental consideration in developing and evaluating tests" (p. 9).

The Standards (AERA et al. 1999) describe validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Zumbo (2009) argues that it is important to make a distinction between validity evidence and the process of validation. He argues that "validity is the explanation for the test score variation, and validation is the process of developing and testing the explanation" (p. 69). He further explains that the "process of validation involves consideration of the statistical methods, as well as the psychological and more qualitative methods of psychometrics, to establish and support the inference to the explanation" (i.e., validity) (Zumbo 2009, p. 70). Validation practices include such methods as: indicators and descriptors of subject matter expert agreement/disagreement over (content) elements, factor analytic and structural equation modeling approaches to internal structure and measurement invariance, reliability and validity coefficients, and response times and descriptions of response option choices. Validation practices can be thought of as the tools that researchers use to build their argument and justification for the test score inference or explanation.

Research syntheses of test validation practices are still relatively new with little agreed-upon methodology. Based on an examination of ten available research syntheses of test validation practices, we found that syntheses may focus on articles in one or more journals or one or more reviews or entries in a single source (such as the *Directory of Unpublished Experimental Mental Measures* or *Mental Measurements Yearbook*). One measure or multiple measures may form the basis of the synthesis. Some authors choose to include all articles, reviews, or entries or a random or systematic selection. Articles, reviews, or entries may come from a single year, cover a range of years, or focus on particular years. None of the syntheses we examined appeared to use search terms to select articles.

Most research syntheses of test validation practices have focused on reporting the frequency that reliability and validity are reported for measures (e.g., Barry et al. 2013; Hogan and Agnello 2004; Meier and Davis 1990; Qualls and Moss 1996; Slaney et al. 2009, 2010; Whittington 1998). Some studies examined whether these frequencies differed for established, new, or modified measures (Barry

[1] Henceforth referred to as *The Standards*.

et al. 2013; Qualls and Moss 1996; Slaney et al. 2009, 2010; Whittington 1998) or by type of measure or journal (Cizek et al. 2008; Qualls and Moss 1996; Slaney et al. 2009, 2010). Many studies examined the types of reliability (e.g., internal consistency, test-retest, alpha) or validity (based on the present sample or previous research, content evidence, construct evidence, internal structure) evidence presented (e.g., Barry et al. 2013; Cizek et al. 2008; Hogan and Agnello 2004; Jonson and Plake 1998; Qualls and Moss 1996; Slaney et al. 2009, 2010). Whittington (1998) examined whether sample characteristics were taken into account when reporting reliability and validity evidence and Slaney et al. (2009, 2010) examined the extent to which researchers followed a logical order in their presentation of reliability and validity evidence. Cizek, Bowen, and Church (2010) narrowed their focus to the frequency that consequences of testing, one of the five sources of validity identified by *The Standards* (AERA et al. 1999), are reported. Several syntheses explicitly focused on the extent that testing practices may be influenced by testing standards (e.g., Jonson and Plake 1998; Qualls and Moss 1996) or validity theory (e.g., Cizek et al. 2008).

The findings of previous research syntheses of test validation practices suggest that (a) the frequency of reporting reliability and validity evidence each seems to have increased generally over time, although this may vary by journal or field of study, (b) reporting of both reliability and validity evidence seems to have increased generally over time but is much less frequent than reporting either type of evidence on its own, (c) there is a failure to take into account sample characteristics when reporting reliability and validity evidence based on previous research, (d) there is mixed evidence as to whether and how the status of a measure (as pre-existing or new/modified) is related to the frequency of presenting reliability and validity evidence, (e) internal consistency estimates of reliability, which almost always consist of Cronbach's alpha, are reported far more frequently than test-retest reliability estimates, (f) validity evidence is often not reported for all measures in a study, tends to be limited in terms of the amount of evidence provided, and is typically poorly reported, (g) some forms of construct validity evidence tend to be reported more often, (h) there is mixed evidence as to the relative frequency of validity evidence such as factor structure and content evidence, (i) validity evidence such as developmental changes, effect of experimental variables, response processes, and consequences of testing is rarely reported, and (j) there is a disconnection between validity theory, test standards, and validation practice.

Only one research synthesis appears to have examined validation practice with a single measure over time (i.e., Jonson and Plake 1998) and the focus of that study was to use MMY reviews over five periods of validity history to examine the relationship between test standards and validation practices. The purpose of the present synthesis was to examine validation practice with a single, well-known, and widely used measure, the *Satisfaction with Life Scale* (SWLS; Diener et al. 1985). Specifically, we aimed to examine a comprehensive list of validation studies of the SWLS to determine the sources of evidence provided and report, in more detail, the kinds of evidence provided for each source, the rationale for steps taken, criteria used, and the logic adopted for the process involved for each procedure. This study

will contribute to the small but growing literature on validation synthesis by (a) exploring validation practice in more detail, and (b) providing a foundation upon which further validation evidence for the SWLS can be built.

Method

Data Source and Collection

We conducted a literature search for articles on the SWLS containing psychometric or validation evidence using the PsycINFO database. Because the SWLS is used in a variety of disciplines and cultural contexts, and has been translated into several languages, PsycINFO was considered the optimal data source. It is the largest resource devoted to peer-reviewed literature in behavioral science and mental health and includes roughly 2,500 international periodicals, publications from more than 50 countries and journals in 20 languages (American Psychological Association 2013). The search history included publications from 1985 (publication date of the SWLS) to July, 2012. A literature search using the search terms "Satisfaction with Life Scale" and "valid*', "reliability", "psychometrics", "factor analysis" "measurement", or "measurement invariance" was used to capture studies whose purpose was to provide validity and reliability evidence for the SWLS. Because 'satisfaction with life' is a general and widely used term, "Satisfaction with Life Scale" was used as a title search term alongside the other terms listed above. Reference sections of identified articles were also used to identify relevant articles. All studies were screened to determine that: (a) the intent of the study was to provide reliability or validity evidence for the SWLS (as opposed to it being used as a comparison measure or assessment tool in differing research contexts), (b) no modified versions (with the exception of translated versions) of the scale were used, and (c) studies were peer-reviewed.

Coding of Studies

We developed a detailed coding sheet to identify and record validation procedures used in each study. The coding sheet was organized according to the sources of validity evidence as outlined in *The Standards* (AERA et al. 1999): (1) test content, (2) internal structure, (3) relations to other variables, (4) response processes, and (5) consequences of testing. As our intention was to provide a detailed account of the reasoning behind the evidence presented, each category was further broken down to document the rationale for steps taken, criteria used, and the logic adopted for the process involved for each procedure. Two additional sections were added to document reliability evidence and translation methods. Although reliability may

not be included in *The Standards* as evidence of validity, it is a necessary condition for validity (Hubley and Zumbo 2011, 2013). Therefore, it is relevant to examine whether a validation study provided any indication of the reliability of the measure's scores within the context specific to the population. Translation methods were also considered given the large number of translated versions of the SWLS that appeared in our search. As each translation is, in essence, a creation of a new measure, it is important that researchers identify the methods used to create the measure. Details regarding the coding of each section are as follows:

Translations and Adaptations of the SWLS. We first identified if the measure was previously translated or newly translated or if use of a translated measure was suspected but not identified (e.g., sample suggested use of a non-English version of the measure but this was not identified in the paper). Where a newly translated measure was used, we coded for the method of translation used, qualifications of the translators, and whether any pilot tests were conducted.

Reliability. We coded for the presence of internal consistency estimates and test-retest reliability estimates based on the present sample. Alternate forms reliability and inter-rater reliability were not recorded as, respectively, there are no alternate forms of the SWLS scale and no rater decisions in scoring the SWLS. When an internal consistency estimate was provided, we noted what estimate was used, and coded for whether a criterion was identified for the estimate presented. Finally, we coded for whether (corrected) item-total correlations and (average) inter-item correlations were reported. When a test-retest reliability estimate was provided, we coded for whether the test-retest interval was reported; if this was the case, we recorded the length of the interval and whether a rationale for the chosen test-retest interval was provided.

Sources of Validity Evidence. Each study was examined to identify and explore what sources of validity evidence, as outlined in *The Standards* (AERA et al. 1999), were provided. Each category was further subdivided as follows:

Test Content. *The Standards* (AERA et al. 1999) dictate that "item selection, response formats, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers" (p. 44). To determine that inferences drawn from test scores are applicable across groups being tested, evidence must be presented to show that the construct being examined is clearly defined, the selected items accurately represent the construct, the process used in generating and evaluating test items is documented, and results of all empirical analyses conducted in the test development and review process have been presented. We descriptively coded to determine if: a) the construct being examined was clearly defined, b) items were generated based on a literature search, other measures of life satisfaction or related constructs (e.g., well-being, quality of life), or feedback from the target population (i.e., experiential experts), c) content experts (e.g., subject matter experts or experiential experts) were consulted to examine elements of the measure, and d) whether any reference was made to item representation (e.g., of different aspects of life satisfaction), construct under-representation, and construct irrelevant variance.

Internal Structure. To demonstrate that the interpretation of a test reflects the construct it proposes to measure, evidence of its internal structure must be presented. Multivariate statistical techniques are used to examine whether "score variability attributable to one dimension was much greater than the score variability attributable to any other dimension scores obtained from one group" (AERA et al. 1999, p. 20). We first reported whether an exploratory factor analysis (EFA), confirmatory factor analysis (CFA), or both analyses were conducted. When an EFA was conducted, we noted the type of EFA used (i.e., principal components analysis (PCA) or true factor analysis (FA)) and coded for the following information: whether criteria were stated a priori for determining the number of factors, the criteria used to determine the number of factors (i.e., eigenvalues > 1, scree plot, parallel analysis, percentage of variance explained), whether factor loadings were reported, whether the criterion (e.g., $>.35$) used for determining if an item loads on a factor was reported, and whether percentage of variance explained was reported. If more than one factor was identified, we also recorded information about the types of rotation methods used. When CFA was conducted, we noted the software used and coded for whether researchers reported the number of factors expected and the items expected to load on each factor (if more than one factor expected), the fit indices used, and the rationale and criteria reported for the chosen fit indices. For those studies that examined measurement invariance, we recorded the software used, the type of invariance examined (e.g., gender invariance), the fit indices used, the criteria reported for the chosen fit indices, and the procedures and rationale for the invariance procedures used.

Relations to Other Variables. When relations to other variables were presented as validity evidence, it is clear in *The Standards* (AERA et al. 1999) that the theoretical rationales behind the selection of those variables and "evidence concerning the constructs represented by the other variables as well as their technical properties, should be presented or cited" (p. 20). Questions regarding the degree of association between the measure being examined (e.g., SWLS) and measures representing similar and dissimilar constructs (i.e., convergent and discriminant measures) must be addressed and shown to be consistent with theoretical expectations. The same is true when quasi-experimental or experimental evidence is presented (e.g., known-group differences based on demographic variables or interventions). When evidence is presented that involves assessing relationships with criterion variables, *The Standards* (AERA et al. 1999) notes that "information about the suitability and technical quality of the criteria should be reported" (p. 21). We recorded (a) the terms that researchers used to describe the validation process (e.g., relations to other variables, construct validity, concurrent validity, convergent validity), (b) whether researchers stated their expectations a priori, (c) the types of measures they included (e.g., discriminant measures) and whether terminology used was incorrect (e.g., confusing criterion evidence with convergent evidence), (d) whether any theoretical or empirical rationale was presented for the measures or variables selected, (e) whether reliability evidence (based on the present sample) was reported for the measures chosen, and (f) how the researchers used the evidence

presented (e.g., magnitude, direction, statistical significance of validity coefficients) to make their conclusions about validity.

Response Processes. Whenever a test involves interpretations that presume underlying psychological or cognitive processes used by respondents, observers, or scorers, empirical evidence in support of those premises should be provided. For example, if the SWLS is meant to involve an overall cognitive judgment of one's life by respondents, then empirical evidence should be provided that such a process is taking place. As approaches to examining response processes can be quite varied and less prescribed than some other sources of validity evidence, we simply described any practices but kept in mind some typical approaches (e.g., think-aloud protocols, cognitive interviewing, completion times, documenting or recording responses to items).

Consequences of Testing. The intended social consequences and unintended side effects (Hubley and Zumbo 2011, 2013) of legitimate test interpretation are "relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components" (AERA et al. 1999, p. 16). When claims are made regarding the benefits of testing beyond the direct interpretation of test scores (e.g., use of a measure will result in reduced costs or more efficient employee selection), evidence is also needed. Like the response processes section, approaches to consequences of testing can be quite varied and less prescribed than some other sources of validity evidence, so we simply described any use of the words "consequences", "consequential validity", "effects of", "impact of", "implications", and "clinical implications".

Results

Our literature search yielded 35 articles that fit the criteria for inclusion in our study. In several cases, the authors conducted multiple studies using different samples within a single article. For example, a single journal article may have included a group of university students to examine internal structure, a different group of university students to examine dimensionality, and a third group using adolescents to examine relations to other variables. In these cases, each study was treated as an independent study and coded accordingly. This resulted in 46 studies. Of those studies, 31 (67.4 %) involved translated versions of the SWLS. In terms of reliability evidence and the broad categories of sources of validity evidence as outlined in *The Standards*, 37 studies (80.4 %) conducted reliability analyses, 39 studies (84.8 %) examined internal structure, 20 studies (43.5 %) examined relations to other variables, and two studies examined response processes (4.3 %). No studies examined test content or consequences of testing (see Table 4.1). Three studies (6.5 %) provided only reliability evidence. Of the 43 studies providing validity evidence, 25 (58.1 %) examined one source of validity evidence and 18 (41.9 %) examined two sources of validity evidence ($M = 1.33$, $SD = 0.60$). A total of 33 studies (71.7 %) examined both reliability and validity evidence.

Table 4.1 Reliability and validity evidence across studies

Article	Language	Sample	Reliability	Sources of validity evidence[a]			Number of validation sources
				Internal structure	Relations to other variables	Response processes	
Diener et al. (1985) (Study 1)	English	Psychology undergraduates in U.S.	x	x			1
Diener et al. (1985) (Study 2)	English	Psychology undergraduates in U.S.			x		1
Diener et al. (1985) (Study 3)	English	Elderly in U.S.			x		1
Arrindell et al. (1991)	Dutch	Adult clinical outpatients in the Netherlands	x	x	x		2
Pavot et al. (1991) (Study 1)	English	Elderly in U.S.	x	x	x		2
Pavot et al. (1991) (Study 2)	English	Undergraduates in U.S.	x	x	x		2
Neto (1993)	Portuguese	Adolescents in Portugal	x	x	x		2
Shevlin and Bunting (1994)	English	Psychology undergraduates in Ireland		x			1
Lewis et al. (1995)	English	Undergraduates in Ireland		x			1
Abdallah (1998)	Arabic	Undergraduates in the West Bank	x	x	x		2
Shevlin et al. (1998)	English	Undergraduates in Britain	x	x			1
Arrindell et al. (1999)	Dutch	Young Dutch community adults	x	x	x		2
Lewis et al. (1999)	Czech	Undergraduates in Czech Republic	x	x			1
Pons et al. (2000)	Spanish	High school students and elderly in Spain		x			1
Atienza et al. (2003)	Spanish	High school students in Spain		x			1
Westaway et al. (2003)	English	Adults in South Africa	x	x	x		2
Vautier et al. (2004)	French	Adults in France	x	x			1
Vittersø et al. (2005)	Norwegian/ Greenlandic	Adults in Norway and Greenland	x			x	1

Navrátil and Lewis (2006)	Czech	Psychology undergraduates in Czech Republic	x				0
Tucker et al. (2006)	Russian/English	Students and community adults in the U.S. and Russia		x			1
Wu and Yao (2006)	Taiwanese	Undergraduates in Taiwan		x			1
Květon et al. (2007)	Czech	Undergraduates in Czech Republic	x				0
Hultell and Gustavsson (2008)	Swedish	Teachers in Sweden	x	x			1
Siedlecki et al. (2008)	English	Community adults in the U.S.	x	x			1
Wu and Wu (2008)	Taiwanese	Community adults with schizophrenia in Taiwan	x	x			1
Wu and Wu (2008) (Study 2)	Taiwanese	Community adults with schizophrenia in Taiwan	x	x	x		2
Gouveia et al. (2009)	Brazilian/ Portuguese	High school students, undergraduates, community members, teachers, and physicians in Brazil	x	x	x		2
Laranjeira (2009) (Study 1)	Portuguese	Students, patients, and health professionals in Portugal	x	x		x	2
Laranjeira (2009) (Study 2)	Portuguese	Patients in Portugal	x				0
Laranjeira (2009) (Study 3)	Portuguese	Students, patients, and health professionals in Portugal			x		1
Slocum-Gori et al. (2009)	English	Community adults in Canada		x			1
Swami and Chamorro-Premuzic (2009)	Malay	Adults in Malaysia	x	x			1
Wu et al. (2009) (Study 1)	Taiwanese	Undergraduates in Taiwan	x	x			1
Wu et al. (2009) (Study 2)	Taiwanese	Adolescent athletes in Taiwan	x	x			1

(continued)

Table 4.1 (continued)

Article	Language	Sample	Reliability	Sources of validity evidence[a]			Number of validation sources
				Internal structure	Relations to other variables	Response processes	
Anaby et al. (2010)	Hebrew	Working adults in Israel	x	x	x		2
Durak et al. (2010) (Study 1)	Turkish	University students in Turkey	x	x	x		2
Durak et al. (2010) (Study 2)	Turkish	Correctional officers in Turkey	x	x	x		2
Durak et al. (2010) (Study 3)	Turkish	Elderly in Turkey	x	x	x		2
Howell et al. (2010) (Study 1)	English	Undergraduates in U.S.	x	x	x		2
Howell et al. (2010) (Study 2)	English	Undergraduates in U.S.	x	x			1
Howell et al. (2010) (Study 3)	English	U.S. undergraduates and adults on social networking sites	x	x	x		2
Bai et al. (2011)	Chinese	Adults in China	x	x			1
Clench-Aas et al. (2011)	Norwegian	Adults in Norway	x	x			1
Glaesmer et al. (2011)	German	Individuals ages 14–91 in Germany	x	x	x		2
Athay (2012)	English	Caregivers of mentally ill youth in the U.S.	x	x			1
Sancho et al. (2012)	Portuguese	Elderly in Southern Africa	x	x	x		2
		Totals:	**36**	**39**	**20**	**2**	**M = 1.33 (SD = 0.60)**

[a]No studies provided evidence based on test content or consequences of testing

Translations and Adaptations of the SWLS

As noted above, 67.4 % of the studies sampled involved translated versions of the SWLS, which includes translations into Arabic, Brazilian-Portuguese, Chinese, Czech, Dutch, French, German, Greenlandic, Hebrew, Malay, Norwegian, Portuguese, Russian, Spanish, Swedish, Taiwanese, and Turkish. These studies come from 25 articles; 12 articles (48.0 %) involved newly translated versions of the SWLS and 9 (32.0 %) used a pre-existing translated version of the scale.[2] In five articles (20 %), no information was provided about the version used; in these cases, we assumed the test was administered in the sample population's dominant language and involved a translated version of the SWLS. Of the 12 articles involving newly translated versions of the SWLS, all but one essentially used forward and backward translation with multiple individuals involved in the translation process. In nearly all cases, very brief descriptions with little elaboration were provided of the translation procedures used. Only one article indicated the authors incorporated a cultural adaptation in their translation process. In six articles (50.0 %), translation guidelines were cited. In 10 articles (83.3 %), it was noted who did the translations but only half provided any information about the translators' qualifications and that was primarily limited to whether they were native speakers, bilingual, or independent translators. In four studies (33.3 %), pilot tests were conducted but little to no information was provided.

Reliability

Thirty-seven (80.4 %) of the 46 studies provided reliability estimates. Of those studies, 33 (89.2 %) provided an internal consistency estimate. The most commonly identified internal consistency estimate was Cronbach's alpha (27 studies; 81.8 %). Five[3] (15.2 %) of the 33 studies provided an "internal consistency coefficient" but were not clear as to which estimate was used. In separate single studies, model-based omega and ordinal alpha were provided in addition to Cronbach's alpha. Finally, another study assessed reliability using parameters estimated from CFA models. In terms of citing criterion values for acceptable internal consistency, only one study clearly cited a criterion (i.e., .70 or higher). Another study made reference to "acceptable" or "satisfactory" alphas of .80 and cited Cronbach's (1951) article, but it is unclear whether a criterion was being listed or the obtained alphas were simply being described.

[2] The total number of articles do not sum to 25 as one article included a new translation in one language and presumably a pre-existing version in another language.

[3] Three studies were contained within a single article wherein the author conducted reliability analyses on three different samples.

Seven (18.9 %) out of 37 studies reported inter-item correlation information; average inter-item correlations were reported in four studies and inter-item correlation tables were provided in another three studies. Twenty (54.1 %) out of 37 studies reported item-total, or corrected item-total, correlations; only five (25 %) of these studies reported an acceptable value (of >.40 or >.50) for evaluating the obtained correlations.

Seven (18.9 %) out of 37 studies examined test-retest reliability, with all studies reporting the time interval between administrations. Intervals examined were 1–2 days, 1 week, 2 weeks, 1 month, 2 months, 3 months, and 6 months. One study examined both 2-week and 1-month intervals and another study examined both 3- and 6-month intervals. Three (42.9 %) of the seven studies provided a rationale for the time interval chosen.

Nine (21.7 %) out of 46 studies did not provide reliability evidence; all but two studies either focused on the internal structure of the SWLS using CFA or examined measurement invariance.

Internal Structure

Thirty-nine (84.8 %) out of 46 studies examined internal structure. Of those 39 studies, 12 studies (30.8 %) conducted exploratory factor analysis, 23 studies (59.0 %) used confirmatory factor analysis, 3 studies (7.7 %) used both methods, and 1 study (2.6 %) was not clear about which approach was used.[4]

Exploratory Factor Analysis. Of 15 studies (i.e., 12 studies using EFA plus 3 studies using both EFA and CFA), 10 studies (66.7 %) used principal components analysis (PCA), 4 studies (26.7 %) used common factor analysis (FA), and 1 study (6.7 %) did not identify the method used. Of the four studies using FA, three studies (75.0 %) used principal axis factoring and one study (25.0 %) used maximum likelihood (ML) as the type of extraction method. No studies stated any criteria a priori for identifying the number of factors. Of the 15 EFA studies, 7 (46.7 %) used 'eigenvalues greater than one' as a criterion, 4 (26.7 %) used scree plots, and 3 (20.0 %) used a combination of both criteria. All studies reported the amount of variance explained by the single factor found, but no studies used a criterion value for the amount of variance explained to decide the number of factors. All but one study (93.3 %) reported factor loadings, but no studies identified a criterion (e.g., >.40) to determine if an item loaded on the factor. No EFA study reported finding more than one factor so other EFA considerations, such as factor rotation, were not explored.

[4] The focus of this methodological article was on describing steps to identify essential unidimensionality that could be used with either EFA or CFA. SWLS data were used as an example. Because it was unclear as to whether the researchers actually used CFA or EFA analyses with this data, this study was not included in the base rate counts in subsequent internal structure sections.

Confirmatory Factor Analysis. Of 26 studies (i.e., 23 studies using CFA plus 3 studies using both EFA and CFA), 24 studies (92.3 %) specified the software program used for analysis. Between 1985 and 2008, LISREL was used predominantly (i.e., in 11 out of 14 studies; 78.6 %). From 2009 through 2012, a more varied array of software programs were used, including Amos, M Plus, EQS, and SAS. Of the 26 CFA studies, 24 (92.3 %) specified the number of factors expected. The studies used between one and eight fit indices ($M = 4.5$, $SD = 1.84$) to evaluate model fit. The most commonly used fit indices were CFI (21/26; 80.8 %), RMSEA (20/26; 76.9 %), χ^2 (16/26; 61.5 %), TLI/NNFI (13/26; 50.0 %), and SRMR (13/26; 50.0 %) (see Table 4.2). Citation of criteria for the range of acceptable values per fit index varied across the 26 studies: 15 studies (57.7 %) provided criteria for all of the indices, 5 studies (19.2 %) provided criteria for some fit indices but not others, and 6 studies (23.1 %) provided no criteria. Only one study (3.8 %) stated the rationale for the fit indices chosen.

Measurement Invariance. A total of 16 (34.8 %) out of 46 studies from 15 articles examined measurement invariance (see Table 4.3). Approximately half (53.3 %) of the articles included a reference to 'invariance' in the title of the article. Of the 16 studies, most ($n = 14$; 87.5 %) involved a SWLS translated into a language other than English. Generally, the description of measurement invariance was similar across all studies in terms of conducting multi-group CFA using ML estimation and covariance matrices. Thirteen studies (81.3 %) indicated the software package used to conduct analyses; prior to 2009, all studies used LISREL. After that, studies used either AMOS or MPlus. The studies used between one and eight fit indices ($M = 4.4$, $SD = 1.82$) to evaluate the fit of invariance models. The most commonly used fit indices were RMSEA (14/16; 87.5 %) and CFI (13/16; 81.3 %), followed by TLI/NNFI (8/16; 50.0 %), SRMR (5/16; 31.3 %), GFI (3/16; 18.8 %), and NFI (2/16; 12.5 %); a few other indices were used only once. Overall, age and gender were each examined in half of the studies. More specifically, out of 16 studies, 2 (12.5 %) examined age invariance alone, 3 (18.8 %) examined gender invariance alone, and 4 (25.0 %) examined both age and gender invariance. The age divisions varied considerably across the studies both in terms of the number of categories (ranging from two to four categories), the age range of the entire sample, and where age cut-offs were made. Four studies (25.0 %) only examined other types of invariance: two studies from one article examined longitudinal invariance (over 2 month intervals) and two studies examined invariance across different samples (e.g., students, correctional officers, and elderly). One study (6.3 %) examined gender invariance and ethnic invariance. Two studies (12.5 %) examined age invariance and other invariance (i.e., invariance across scattered versus successive item order, nationality). Finally, one study (6.3 %) examined age, gender, education, income, and residence (i.e., metropolitan, town, or rural) invariance.

Table 4.2 Software program and fit indices used in confirmatory factor analysis studies

	Software	χ^2	χ^2/df	GFI	PGFI	AGFI	PNFI	IFI	NFI	TLI/ NNFI	RMSEA	SRMR	RMSR	CFI	AIC	CAIC	RMR	CN	Total per study
Shevlin and Bunting (1994)	LISREL	x		x		x							x						**4**
Lewis et al. (1995)	LISREL	x		x		x							x						**4**
Shevlin et al. (1998)	LISREL	x																	**1**
Lewis et al. (1999)	LISREL	x		x					x										**3**
Pons et al. (2000)	LISREL		x											x					**2**
Atienza et al. (2003)	LISREL		x								x			x					**3**
Vautier et al. (2004)	EQS	x									x	x		x					**4**
Tucker et al. (2006)	–			x						x	x								**3**
Wu and Yao (2006)	LISREL	x							x	x	x			x	x	x	x		**8**
Hultell and Gustavsson (2008)	LISREL										x	x		x					**3**
Siedlecki et al. (2008)	–	x								x	x			x					**4**
Wu and Wu (2008) (Study 1)	LISREL	x								x	x	x		x					**5**
Wu and Wu (2008) (Study 2)	LISREL	x								x	x	x		x					**5**
Gouveia et al. (2009)	LISREL		x	x						x	x	x		x					**6**
Swami and Chamorro-Premuzic (2009)	AMOS	x		x	x						x			x	x		x	x	**8**

Wu et al. (2009) (Study 1)	MPlus	x								x	x	x		x						**5**
Wu et al. (2009) (Study 2)	MPlus	x								x	x	x		x						**5**
Anaby et al. (2010)	EQS										x			x						**2**
Durak et al. (2010) (Study 1)	AMOS	x	x					x		x	x	x		x						**7**
Durak et al. (2010) (Study 2)	AMOS	x	x					x		x	x	x		x						**7**
Durak et al. (2010) (Study 3)	AMOS	x	x					x		x	x	x		x						**7**
Bai et al. (2011)	MPlus									x	x	x		x						**4**
Clench-Aas et al. (2011)	AMOS				x		x				x			x						**4**
Glaesmer et al. (2011)	AMOS			x					x	x	x			x						**5**
Athay (2012)	SAS			x								x		x						**3**
Sancho et al. (2012)	EQS	x		x							x	x		x						**5**
Total per indices:		**16**	**6**	**9**	**2**	**2**	**1**	**3**	**3**	**13**	**20**	**13**	**2**	**21**	**2**	**1**	**2**	**1**		

GFI Goodness of Fit Index, *PGFI* Parsimony Goodness-of-Fit Indicator, *AGFI* Adjusted Goodness of Fit Index, *PNFI* Parsimony Normed Fit Index, *IFI* Incremental Fit Index, *NFI* Normed Fit Index, *TLI* Tucker Lewis Index also known as *NNFI* Non-normed Fit Index, *RMSEA* Root Mean Squared Error of Approximation, *SRMR* Standardized Root Mean Square Residual, *RMSR* Root Mean Square Residual, *CFI* Comparative Fit Index, *AIC* Akaike's Information Criterion, *CAIC* Consistent Akaike Information Criterion, *RMR* Root Mean Square Residual, *CN* Hoelter's Critical N

– = not identified, χ^2 = Chi Square, χ^2/df = Chi-Square change/Degrees of freedom

Table 4.3 Measurement invariance studies

Article	Language	Sample	Invariance mentioned in title	Type of invariance			Software used
				Age	Gender	Other	
Shevlin et al. (1998)	English	Undergraduates in Britain	x		x		LISREL
Pons et al. (2000)	Spanish	High school students and elderly in Spain	x	x			LISREL
Atienza et al. (2003)	Spanish	High school students in Spain	x		x		LISREL
Vautier et al. (2004)	French	Adults in France		x		x	–
Tucker et al. (2006)	Russian/English	Students and community adults in the U.S. and Russia	x	x		x	–
Wu and Yao (2006)	Taiwanese	Undergraduates in Taiwan	x		x		LISREL
Hultell and Gustavsson (2008)	Swedish	Teachers in Sweden		x	x		LISREL
Siedlecki et al. (2008)	English	Community adults in the U.S.		x			–
Gouveia et al. (2009)	Brazilian/ Portuguese	High school students, undergraduates, community members, teachers, and physicians in Brazil				x	LISREL
Swami and Chamorro-Premuzic (2009)	Malay	Adults in Malaysia			x	x	AMOS
Wu et al. (2009) (Study 1)	Taiwanese	Undergraduates in Taiwan	x			x	MPlus
Wu et al. (2009) (Study 2)	Taiwanese	Adolescent athletes in Taiwan	x			x	MPlus
Durak et al. (2010) (Study 3)	Turkish	Elderly in Turkey				x	AMOS
Bai et al. (2011)	Chinese	Adults in China		x	x	x	MPlus
Clench-Aas et al. (2011)	Norwegian	Adults in Norway	x	x	x		AMOS
Glaesmer et al. (2011)	German	Individuals ages 14–91 in Germany		x	x		AMOS
		Totals:	**8**	**8**	**8**	**8**	

– = not identified

Relations to Other Variables

Twenty studies (43.5 %) out of the 46 studies included in this synthesis examined relations to other variables (see Table 4.4). The vast majority (19; 95.0 %) of these studies examined convergent evidence, with far fewer including evidence using discriminant measures, although it should be noted that many studies never directly addressed what measures qualified as convergent or discriminant. One study provided what might be best referred to as known-groups evidence only but three of the other studies also attempted to provide known-groups type evidence (4 of 20 studies; 20.0 %).

With respect to the 19 studies examining convergent and/or discriminant evidence, the total number of demographic variables or measures used per study ranged from 2 to 19 (M = 7.16, SD = 5.09). *The Standards* (AERA et al. 1999) state that, when comparisons with other variables are presented as validity evidence, the rationale behind the selection of those variables and "evidence concerning the constructs represented by the other variables. . .should be presented or cited" (AERA et al. 1999, p. 20). This means that researchers need to clearly state the rationale for both the construct selected and any variables used to represent that construct. Regarding a rationale for *constructs* used, 13 (68.4 %) out of the 19 studies provided no rationale and 6 studies (31.6 %) provided some rationale. When the rationale for constructs was not explicitly stated, it was often implied, because the constructs were used in previous research or comprised some aspect of the construct of subjective well-being (SWB); thus, authors may have thought that explicitly stating a rationale would be redundant. Regarding a rationale for *measures* used, no studies provided a clear rationale as to why they selected the specific measures chosen.

The terminology used to describe convergent and/or discriminant validity evidence varied considerably both across and within the 19 studies and was, at times, incorrect. Only one study avoided using any terms and three studies referred only to "construct validity". Otherwise, the following terms were used: "construct validity" (7 studies), "convergent validity" (6 studies), "discriminant validity" (4 studies), "divergent validity" (1 study), "criterion"(-related) validity (5 studies), "criterial validity" (1 study), "concurrent validity" (4 studies), and "predictive" validity or relationships (3 studies).

When it came to stating in advance the expected relationships among variables, only 2 (10.5 %) of the 19 studies clearly identified what they expected to find and 11 studies (57.9 %) were vague in that the expected findings were not explicitly stated by researchers but it was implied they were based on findings in previous literature; 6 studies (31.6 %) did not indicate any expected findings.

An important piece of information in understanding validity coefficients is to know not only reliability estimates for the SWLS scores but also for the other measures used. Out of 19 studies, 6 (31.6 %) provided reliability estimates for all

Table 4.4 Types of relations with other variables evidence

Authors	Language	Sample	Convergent	Discriminant	Known-groups
Diener et al. (1985) (Study 2)	English	Psychology undergraduates in U.S.	x		
Diener et al. (1985) (Study 3)	English	Elderly in U.S.	x		
Arrindell et al. (1991)	Dutch	Adult clinical outpatients in the Netherlands	x	x	
Pavot et al. (1991) (Study 1)	English	Elderly in U.S.	x		
Pavot et al. (1991) (Study 2)	English	Undergraduates in U.S.	x		
Neto (1993)	Portuguese	Adolescents in Portugal	x		x
Abdallah (1998)	Arabic	Undergraduates in the West Bank	x	x*	
Arrindell et al. (1999)	Dutch	young Dutch community adults	x	x	x
Westaway et al. (2003)	English	Adults in South Africa	x	x*	
Wu and Wu (2008) (Study 2)	Taiwanese	Community adults with schizophrenia in Taiwan	x		x
Gouveia et al. (2009)	Brazilian/ Portuguese	High school students, undergraduates, community members, teachers, and physicians in Brazil	x		
Laranjeira (2009) (Study 3)	Portuguese	Students, patients, and health professionals in Portugal			x
Anaby et al. (2010)	Hebrew	Working adults in Israel	x		
Durak et al. (2010) (Study 1)	Turkish	University students in Turkey	x	x	

(continued)

Table 4.4 (continued)

Authors	Language	Sample	Convergent	Discriminant	Known-groups
Durak et al. (2010) (Study 2)	Turkish	Correctional officers in Turkey	x		
Durak et al. (2010) (Study 3)	Turkish	Elderly in Turkey	x		
Howell et al. (2010) (Study 1)	English	Undergraduates in U.S.	x		
Howell et al. (2010) (Study 3)	English	U.S. undergraduates and adults on social networking sites	x		
Glaesmer et al. (2011)	German	Individuals ages 14–91 in Germany	x		
Sancho et al. (2012)	Portuguese	Elderly in Southern Africa	x		

* = possibly, but not entirely clear

measures used, 1 study (5.3 %) provided reliability estimates for some measures used but not others, and 11 (57.9 %) provided no reliability evidence for the other measures. Finally, one study (5.3 %) provided reliability estimates but it was unclear whether estimates were based on the study sample or previous research.

It was often unclear what information researchers relied on when it came to interpreting the correlations providing convergent and/or discriminant validity evidence. In some cases (6 studies; 31.6 %), it was not mentioned what information was used to interpret correlations; in other cases, some reference was made to statistical significance (7 studies; 36.8 %), sign of the correlation (positive/negative) (9 studies; 47.4 %), magnitude of the correlation (8 studies; 42.1 %), and/or effect size (1 study; 5.3 %).

Four studies attempted to provide validity evidence by comparing groups in a way that is akin to known-groups validity, although none of the studies used this term; rather, researchers referred to this as either construct or discriminant validity evidence. With the exception of additional procedures to complement convergent/discriminant evidence (e.g., use of factor analysis, partial correlations, or multiple regression to examine contributions of different variables to SWLS scores), no other forms of evidence under the heading of 'relations to other variables' were examined.

Response Processes

Two studies provided some evidence related to response processes. In one study, mean time to complete the SWLS as well as ease of use reported by study participants and interviewers was recorded. In another study, a mixed Rasch model was used to identify four latent classes of respondents to the SWLS. The classes tend to reflect differences in the use of the response categories or extreme scores, the difficulty or discriminability of items, or the level of life satisfaction.

Discussion

The purpose of this study was to contribute to the small but growing literature on validation synthesis by (a) exploring validation practices in more detail, and (b) providing a foundation upon which further validation evidence for the SWLS can be built. Thus, our intentions are aimed at measurement and validation specialists, researchers interested in using the SWLS and further examining the validity of inferences made from it, and those individuals who use measures and desire to better understand the validation procedures used to support the inferences drawn from test scores.

While others may have used *The Standards* and the five sources of validation evidence as an inspiration or guide for conducting validation synthesis in the past (e.g., Cizek et al. 2008; Hogan and Agnello 2004; Jonson and Plake 1998), the detailed documentation of procedures and rationales involved in validation practices provided in this study appears to be the first of its kind. If the validation process "begins at the construct definition stage before items are written or a measure is selected, continues through item analysis (even if one is adopting a known measure), and needs to continue when the measure is in use" (Zumbo 1999, p. 11), then a detailed account over time of procedures used, specific to a given test, and within the areas outlined by *The Standards,* is needed.

It is important to note that the majority of studies (67.4 %) included in this synthesis involved translated versions of the SWLS and nearly 50 % of studies involved newly translated versions. Generally, the process that was used to create these translated versions is not well reported. Little information is provided about the individuals who conducted the translation and there is relatively little use of pilot testing reported. Previous research syntheses of validation practices provide no explicit discussion of translated versions of measures, although these measures may be included under 'modified measures' in some studies. Translated versions of the SWLS are essentially new measures and so it is critical that the process used to translate the measures is well documented.

Reliability Evidence

Reliability evidence for SWLS scores was consistently well documented across studies. Internal consistency was examined most often (89.2 % of studies). The internal consistency estimate most commonly used was Cronbach's alpha (81.8 % of the time), which shows that classical test theory approaches to reliability still dominate, at least with respect to the SWLS. No study clearly stated a criterion for an acceptable reliability estimate, although it may be viewed as common knowledge to expect estimates of .80 or higher.

In several cases, researchers reported (average) inter-item correlations or (corrected) item-total correlations but failed to identify or discuss acceptable values for, the role of, or how to interpret, these correlations. Inter-item correlations indicate the degree to which items correlate with one another. They are particularly useful in item and test construction to identify whether an item correlates poorly with other items in a test, or whether an item correlates strongly with some items but not others. Both patterns suggest that one may be tapping into another construct altogether (construct irrelevant variance) or that some items tap into another aspect of the construct that the other items are not tapping into (either construct irrelevant variance or construct underrepresentation). Three studies presented inter-item correlations in a table and three studies provided average inter-item correlations. All concluded their results were acceptable, but none discussed the relationship of these correlations to internal consistency or indicated what constitutes an acceptable value, despite the availability of such guidelines. For example, Clark and Watson (1995, p. 316) suggest that, for higher order constructs (such as the SWLS), a mean correlation of .15 to .20 is acceptable whereas for constructs that are more narrowly defined (e.g., talkativeness), a higher mean inter-correlation (i.e., .40 to .50) would be needed. It has been suggested by others (e.g., Netemeyer et al. 2003; Clark and Watson 1995) that the little attention paid to inter-item/average inter-item correlations may be problematic, and that the average inter-item correlation provides a more useful index of internal consistency than does coefficient alpha, the predominant estimate reported in the studies examined. Because coefficient alpha is a function of the number of items in a test and the average inter-correlation among test items, it is possible to achieve a high internal consistency reliability estimate by: (a) having a large number of items, (b) having items that are highly correlated, or (c) a combination of the two. Similarly, Cortina (1993) suggests that coefficient alpha is problematic for scales with more than 40 items. In such cases, the coefficient alpha value may be driven more by the number of items than the magnitude of the correlations among items. The result can be a high internal consistency estimate for a test with items that may correlate rather poorly with one another. Having said this, the small number of items comprising the SWLS limits their influence on the value of coefficient alpha. Thus, alpha will, in this case, be driven more by the magnitudes of the inter-item correlations and is arguably an adequate and more straightforward indicator of internal consistency. Still, more

attention should be paid to inter-item correlations or average inter-item correlations and relaying to the reader what values are acceptable.

The other problematic area of reporting with respect to reliability involved (corrected) item-total correlations. Item-total correlations are computed by correlating the score for a single item with the total score on a scale, and corrected-item total correlations are computed by correlating the score of a single item with the total score on a scale based on the remainder of the items. Researchers should provide some indication of what values are considered acceptable to aid in interpreting the results presented. As a general rule, low or near zero correlations indicate problematic items (Hubley and Zumbo 2011). Generally, values of .50 and above are found to be acceptable values (Netemeyer et al. 2003). It is valuable when both (corrected) item-total correlations and inter-item correlations are presented. One can think of (corrected) item-total correlations as a photograph and inter-item correlations as a sort of zoom lens allowing for a more detailed examination of the items in question. In the case of the SWLS, few studies provided either of these values, and no studies provided both.

Not surprisingly given evidence from previous syntheses, test-retest reliability estimates were provided less often (18.9 % of studies). These studies all reported the test-retest interval but, in a majority of cases, did not provide a rationale for the length of interval chosen. This rationale is an important element needed to assess the obtained estimate because the interval needs to make sense given the expected stability of the construct. With the SWLS, it would be important to choose a time interval length not so short that respondents might recall their responses to items but also not so long that one might anticipate changes to occur in their satisfaction with life. Put another way, it is critical to be able to assume that respondents are not simply trying to report their previous responses and that no real change in satisfaction with life has occurred in order to appropriately evaluate a given test-retest reliability coefficient.

Sources of Validity Evidence

In terms of the five sources of validity evidence as outlined in *The Standards*, only three sources of evidence have been presented for the SWLS. The two primary sources consisted of internal structure and relations to other variables; two studies examined response processes. No studies examined evidence based on test content or consequences of testing.

Internal Structure. Internal structure is the most common type (84.8 % of studies) of validity evidence examined for the SWLS. The majority (59.0 %) of the studies examining internal structure used CFA. The number of factors expected, fit indices used, and software used for analysis were, overall, well reported. The number of fit indices used per study ranged from one to eight, with less than five fit indices used on average. Information needed, but lacking, involves the rationale for fit indices chosen, and, in some cases, criterion values for the fit indices chosen.

When conducting CFA, a rationale for the fit indices used should be provided. Once a model is chosen and estimated, the "fit" of the model must be determined. The fit of a model is largely influenced by sample size and assumptions regarding score distributions and independence assumptions (Tabachnick and Fidell 2013). Although there are a number of indices from which to choose, as a general rule, consistency in results across indices indicate a good fitting model (Tabachnick and Fidell 2013). However, because what fit indices you use influence the results obtained, it is informative to report a rationale for those indices. Tabachnik and Fiddell note that, "numerous measures of model fit have been proposed. In fact, this is a lively area of research with new indices seemingly developed daily" (p. 720). To provide a rationale for the selected fit indices not only indicates that the researcher has considered the influence of details specific to the sample being examined, it also provides a context for other researchers using or developing new indices.

Fewer, but still a significant number of, studies (30.8 %) used EFA. Of these, the predominant method used was principal components analysis (PCA; 66.7 %) rather than common factor analysis (FA; 26.7 %). There appeared to be no association between the time (e.g., in which decade) a study was conducted and the EFA method used. All of the EFA studies conducted found evidence to support a one-factor model. Eigenvalues greater than one (46.7 %) were most commonly used to identify the number of factors, followed by scree plots (26.7 %); few studies (20.0 %) used both criteria. One currently recommended criterion is to use loadings obtained from a parallel analysis as a standard against which obtained loading values can be compared (Hayton et al. 2004). Specifically, this procedure involves comparing the eigenvalues found against those eigenvalues that would be obtained from random numbers generated from a data set that is equivalent in sample size and consists of the same number of variables (Ledesma and Valero-Mora 2007). If the eigenvalues obtained exceed those that are randomly generated, then those components can be retained. None of the SWLS studies used this criterion. All but one study reported factor loadings. Surprisingly, no study appeared to use a criterion (e.g., factor loading $>.40$) to determine if an item loaded on a factor. As well, all studies reported the amount of variance explained by the single factor found, but no studies used this as a criterion value to decide the number of factors. For example, no one explicitly stated that a given factor must explain a minimum of 25 % of the variance explained in order for a factor to be retained or considered worthwhile. Given the small number of items on the SWLS, it probably makes more sense to use CFA and test the fit of a unidimensional structure in future studies. If, however, EFA is used, greater attention needs to be paid to the criteria used for (a) identifying whether items load on a factor and (b) the number of factors.

Measurement invariance of the SWLS across groups was examined in 34.8 % of studies, with most studies focusing on invariance across sex or age groups. Notably, however, 87.5 % of these studies examined invariance of a non-English language version of the SWLS so there is a relative gap in the literature on invariance studies with the English version of the SWLS. It would not be surprising for researchers to want to make comparisons in SWLS scores among different groups (e.g., to

examine sex, age, socio-economic status, ethnic or country differences). However, even if there is other validity evidence to support the inferences made from SWLS scores in the different groups, this does not guarantee that the SWLS functions the same way across groups as required for comparison purposes (Horn and McArdle 1992). Only through evidence for measurement invariance can SWLS total scores be deemed to measure the same attribute across groups. If no evidence is presented to support an adequate level of measurement invariance, any differences found among groups cannot be interpreted unambiguously. As Horn (1991, p. 119) has argued, "Without evidence of measurement invariance, the conclusions of a study must be weak". Thus, if researchers want to compare life satisfaction levels among different groups using the SWLS, evidence of strong or scalar levels of measurement invariance must be shown for SWLS total scores among those groups.

Relations to Other Variables. Validity evidence based on relationships to other variables describes the extent to which there is a relationship between SWLS scores and other variables (whether demographic variables or scores from measures or other variables). Just under half of the studies (43.5 %) addressed relations to other variables; 95 % of studies examined convergent evidence. Many studies never directly addressed what measures qualified as convergent or discriminant, which often made it difficult or impossible to determine whether researchers included, or intended to include, discriminant measures. Moreover, there appeared to be considerable confusion and inconsistency across, or even within, studies as to the appropriate terms to use to describe evidence. Most commonly, criterion-related validity terms (including concurrent or predictive validity) were used to describe convergent evidence. There were three issues related to evidence based on relations to other variables that stood out in this validation synthesis: (a) lack of a clearly state rationale for the selection of constructs and variables, (b) lack of clarity in terms of precisely what researchers expected to find, and (c) poor evaluation of the obtained evidence. We will describe each of these issues in turn.

A clearly stated rationale for why constructs and variables were chosen is generally missing or, at best, very unclear. *The Standards* (AERA et al. 1999) state that, when comparisons with other variables are presented as validity evidence, the rationale behind the selection of those variables and "evidence concerning the constructs represented by the other variables. . .should be presented or cited" (p. 20). For example, if examining the relationship between scores on the SWLS and neuroticism, one needs to provide a rationale for why the construct of neuroticism is being used as well as state a rationale for the specific measure of neuroticism chosen (e.g., the Big Five Inventory subscale of neuroticism). When comparing measures representing the same construct (e.g., life satisfaction or even subjective well-being), there seems to be little point in providing a rationale for why that construct has been selected. However, a rationale for the variable(s) used to measure the construct is needed (e.g., why was a particular single-item measure of life satisfaction chosen for use as opposed to another measure of life satisfaction?). In the case of demographic variables, it is less clear whether a rationale is needed for why researchers have assigned the numbers the way they did. On the one hand, because gender, for example, tends to be clearly defined, it may not be necessary to

justify the variable once you have justified the construct. On the other hand, a variable such as age can have numbers assigned in many different ways (e.g., 1 = 20–49 years (young), 2 = 50+ years (old) vs. 1 = 20–49 years (young), 2 = 50–69 years (middle aged), and 3 = 70+ years (old)). Where the assigning of numbers can alter the construct being examined, the decision about how to categorize the variable may require justification (e.g., why is old = 50+ years in one case vs. 70+ years in another case?).

The Standards (AERA et al. 1999) noted that "when validity evidence includes empirical analyses of test responses together with data on other variables, the rationale for selecting the additional variables should be provided" (p. 20). However, *The Standards* do not explicitly articulate or provide a detailed explanation as to what constitutes a rationale. It is noted that the relationships between scores on the variable of interest and other variables "should be consistent with theoretical expectations" (AERA et al. 1999, p. 20). It is also noted that these variables "might include intended measures of the same construct or of different constructs" (AERA et al. 1999, p. 21). This implies that the rationale requires some theoretical explanation to support why the selected variable (or construct) should or should not be related to the variable (or construct) of interest. Alternatively, or in addition, the rationale could include consistently found empirical evidence of a relationship between the variable of interest and other variables.

The constructs most often used for comparison with the SWLS were subjective well-being (SWB) – including positive and negative affect, personality (particularly neuroticism and extroversion), and psychological constructs (e.g., self-esteem, depressiveness). Of these constructs, SWB was clearly and consistently defined, possibly because the definition is inherent when describing what the SWLS is designed to measure. Most researchers provided a rationale by virtue of explaining how the SWLS is designed to measure the cognitive aspect of life satisfaction. In further situating life satisfaction within SWB, the construct of SWB was fairly well described. Other constructs, such as psychological functioning, perceived health, personality traits, and mental health constructs such as depression and self-esteem, were commonly used but the rationale provided for their use was not clearly articulated. This leaves the reader to wonder why those constructs were chosen, and, by extension, if the researchers themselves had a clear reason for choosing them. Some researchers made mention of relationships to variables without discussing the constructs those variables were designed to capture.

The argument in support of the use of constructs is distinct from the rationale used in support of the variables representing those constructs. *The Standards* (AERA et al. 1999) state that "evidence concerning the constructs represented by the other variables as well as their technical properties, should be presented or cited" (AERA et al. 1999, p. 20). To demand that empirical evidence in support of every variable (measure) chosen be presented may be unmanageable due to page or word restrictions dictated by journals and their editors or place an unreasonable burden on researchers. As well, such information may overwhelm rather than inform the reader. However, some indication as to why the variable was chosen and what construct it was intended to represent is needed. Without some logic to

orient the reader as to where constructs and variables fit within existing literature and a nomological network for the construct and measure of interest, and without the distinction between the two clearly articulated, constructs risk being inconsistently defined. Measures are designed to capture specifically defined constructs. If the definition of the construct varies (or remains undefined) across multiple studies, then the validity of the specific inferences made from the variables (measures) cannot be determined and comparisons across studies cannot be evaluated. As well, information regarding the ability of a measure to consistently capture the intended construct is also compromised. The demographic variables used in the studies examined in this validation synthesis included sex, age, marital status, educational level, employment status, monthly income, health insurance, and sociocultural level. It is important to know how and why researchers constructed the variable(s) as they did to determine comparability across studies. In the studies examined here, the distinction between construct and variable was often blurred, making it difficult to discern arguments in support of a rationale for constructs from those in support of a variable.

In addition to providing a rationale for researchers' choice of constructs/variables, hypotheses or a description of how variables are expected to be related to SWLS scores should be provided based on theory or previous empirical research. Ideally, convergent and discriminant evidence should both be included in a study and these results should be interpreted in relation to each other. Hypotheses or description of expectations should provide information about both the direction and relative magnitude of the expected relationship between the scores and should be stated in advance of the analyses so the obtained evidence may be properly evaluated as supportive of the intended SWLS interpretations or not. Just as statistical procedures used in other areas of evidence (reliability estimates, factor loadings for internal structure) demand criterion values as a means to interpret results obtained, relations to other variables also demands some criterion as a means to interpret the obtained correlations. In essence, researchers provide their own criterion by stating a priori the relationships they expect to find. Without clearly stating this expectation, one is left with a series of correlations of varying magnitudes but no context in which to interpret the immediate study results, their relative standing in relation to a proposed theory, or to the results found in other studies examining similar variables. In the absence of expected values for interpretation, there is no link between results obtained and conclusions drawn.

The problem of not providing hypotheses or a description of how variables are expected to be related to SWLS scores is directly related to the final issue of researchers often presenting either a vague, or lack of, evaluation of the obtained evidence. It was often unclear what information researchers relied on when it came to interpreting the correlations providing convergent and/or discriminant validity evidence. There is a strong tendency for researchers to present correlations without interpreting some, or even all, of them and to then to assert that the findings support validity. In many cases, only a vague reference is made to the sign (positive/negative), magnitude, or statistical significance of the correlations. It is fairly

obvious when reviewing this evidence that researchers do not appear to have a clear sense of how to properly and thoroughly evaluate this kind of evidence.

In a few studies, researchers attempted to provide validity evidence by comparing groups on SWLS scores. Typically, researchers referred to this as either construct or discriminant validity evidence, although it might be better described as known-groups validity. It is notable that in the two studies using demographic variables to form groups, the researchers referred to differences found in previous research whereas in the two studies using pre-existing groups (e.g., clinical sample vs. community members), the differences were presumed rather than based on known differences from previous research or theory. The evidence presented was not particularly strong primarily because the foundation of a previously known difference was not firmly in place. It is also worth noting that it was sometimes a challenge to distinguish between cases in which group differences were simply being examined and those cases in which validity evidence was being presented.

Response Processes. The SWLS is intended to capture the judgmental component of life satisfaction (Diener et al. 1985). When there is a presumption that respondents are using an underlying psychological or cognitive process when responding to test items, *The Standards* recommends that "empirical evidence in support of those premises should be provided" (AERA et al. 1999, p. 20). Of the two studies that addressed response processes, one of those simply examined the mean time to complete the SWLS; while this is useful information, it does not contribute much to our understanding of the underlying process used when responding to the SWLS. The second study used a Rasch model to identify four latent classes of respondents to the SWLS. Further research is needed to better understand the processes used to respond to SWLS items by different groups. For an excellent example of exploring responses processes, readers are referred to Gaderman et al.'s (2011) examination of how children respond to the Satisfaction with Life Scale Adapted for Children (SWLS-C).

In summary, the findings of this synthesis of test validation practices suggest that (a) most psychometric studies on the SWLS are based on a wide range of non-English versions of the SWLS but little attention is paid to the language version of the SWLS or the sample characteristics when reporting reliability and validity evidence from previous research, (b) internal consistency (especially alpha) estimates of reliability are reported far more frequently than test-retest reliability estimates or other indicators such as (average) inter-item correlations or (corrected) item-total correlations, (c) sources of validity evidence for inferences made from the SWLS rely heavily on internal structure and relations to other variables with no evidence presented based on test content or consequences of testing, (d) relations to other variables evidence relies heavily on convergent evidence, and (e) validity evidence in the form of relations to other variables tends to be poorly evaluated and reported. Moreover, as has been reported in previous validation synthesis research, there continues to be a disconnection between validity theory, test standards, and validation practice. While some researchers assert that the lack of some sources of validity evidence reflects either a misunderstanding of the procedures required to demonstrate the quality of evidence presented or disagreement as to what

constitutes validation evidence (e.g., Cizek et al. 2008; Hubley and Zumbo 2013), it seems that much is to be gained by presenting more user-friendly guides to conducting and reporting different sources of evidence and different validation procedures, including a more user-friendly version of *The Standards* (AERA et al. 1999).

Strengths and Limitations

The first strength of this validation synthesis study is that we sought to examine a comprehensive set of peer-reviewed and published validation studies regarding the SWLS. Where other validation synthesis studies have used a random sample of studies to examine validation procedures, we sought to examine all published validation studies found in PsycInfo, one of the largest resources of peer-reviewed literature in behavioral science and mental health. Second, we sought to ground our analysis in procedures proposed by *The Standards*, a widely accepted resource for validation procedures. Third, where many validation synthesis studies have coded according to broad areas of evidence, we sought to examine, in detail, each source of validity evidence found in *The Standards* to identify the specific methods and procedures that researchers use in the process of validating inferences made from the SWLS.

Despite these strengths, there are a number of limitations that affect this study. First, the selected search criteria ruled out any studies that did not use our search terms or those studies where researchers implicitly intended to conduct a validation study but did not explicitly identify their study as such. A second limitation is that the level of detail addressed required a fairly high level of understanding of measurement and statistical methods used and thus some subjective judgment in the coding of evidence. This was particularly evident when coding information related to 'relations to other variables' and reflected the confusion and inconsistency in the terminology used, the lack of a clear framework presented by researchers, and poor evaluation and discussion of this type of evidence.

References[5]

*Abdallah, T. (1998). The Satisfaction with Life Scale (SWLS): Psychometric properties in an Arabic-speaking sample. *International Journal of Adolescence and Youth, 7*, 113–119.

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.

[5] * indicates articles included in the validation synthesis

American Psychological Association. (2013). *PsycINFO Quick Facts*. Retrieved May 14, 2013 from http://www.apa.org/pubs/databases/psycinfo/index.aspx

*Anaby, D., Jarus, T., & Zumbo, B. D. (2010). Psychometric evaluation of the Hebrew language version of the Satisfaction with Life Scale. *Social Indicators Researcher, 96*, 267–274. doi:10.1007/s11205-009-9476-z.

*Arrindell, W. A., Meeuwesen, L., & Huyse, F. J. (1991). The Satisfaction with Life Scale (SWLS): Psychometric properties in a non-psychiatric medical outpatients sample. *Personality and Individual Differences, 12*, 117–123.

*Arrindell, W. A., Heesink, J., & Feij, J. A. (1999). The Satisfaction with Life Scale (SWLS): Appraisal with 1700 healthy young adults in The Netherlands. *Personality and Individual Differences, 26*, 815–826.

*Athay, M. M. (2012). Satisfaction with Life Scale (SWLS) in caregivers of clinically-referred youth: Psychometric properties and mediation analysis. *Administration and Policy in Mental Health and Mental Health Services Research, 39*, 41–50. doi:10.1007/s10488-011-0390-8.

*Atienza, F. L., Balaguer, I., & García-Merita, M. L. (2003). Satisfaction with Life Scale: Analysis of factorial invariance across sexes. *Personality and Individual Differences, 35*, 1255–1260.

*Bai, X., Wu, C., Zheng, R., & Ren, X. (2011). The psychometric evaluation of the Satisfaction with Life Scale using a nationally representative sample of China. *Journal of Happiness Studies, 12*, 183–197. doi:10.1007/s10902-010-9186-x.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2013). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior* [online]. doi:10.1177/1090198113483139

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. doi:10.1177/0013164407310130.

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. doi:10.1177/0013164410379323.

Clark, A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319.

*Clench-Aas, J., Nes, R. B., Dalgard, O. S., & Aarø, L. E. (2011). Dimensionality and measurement invariance in the Satisfaction with Life Scale in Norway. *Quality of Life Research, 20*, 1307–1317.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step by step approach* (3rd ed.). Thousand Oaks: Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

*Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment, 49*, 71–75.

*Durak, M., Senol-Durak, E., & Gencoz, T. (2010). Psychometric properties of the Satisfaction with Life Scale among Turkish university students, correctional officers, and elderly adults. *Social Indicators Research, 99*, 413–429. doi:10.1007/s11205-010-9589-4.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction of Life Scale adapted for children: A focus on cognitive processes. *Social Indicators Research, 100*, 37–60. doi:10.1007/s11205-010-9603-x.

*Glaesmer, H., Grande, G., Braehler, E., & Roth, M. (2011). The German version of the Satisfaction with Life Scale (SWLS): Psychometric properties, validity, and population-based norms. *European Journal of Psychological Assessment, 27*, 127–132.

*Gouveia, V. V., Milfont, T. L., Nunes da Fonseca, P., de Miranda, P., & Coelho, J. A. (2009). Life satisfaction in Brazil: Testing the psychometric properties of the Satisfaction with Life Scale (SWLS) in five Brazilian samples. *Social Indicators Research, 90*, 267–277.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*, 191–205.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*, 802–812. doi:10.1177/0013164404264120.

Horn, J. L. (1991). Discussion of the issues of factorial invariance. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 114–125). Washington, DC: American Psychological Association.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Journal of Experimental Aging Research, 18*, 117–144.

*Howell, R. T., Rodzon, K. S., Kurai, M., & Sanchez, A. H. (2010). A validation of well-being and happiness surveys for administration via the Internet. *Behavior Research Methods, 42*, 775–784. doi:10.3758/BRM.42.3.775.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219–230. doi:10.1007/s11205-011-9843.4.

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.

*Hultell, D., & Gustavsson, J. P. (2008). A psychometric evaluation of the Satisfaction with Life Scale in a Swedish nationwide sample of university students. *Personality and Individual Differences, 44*, 1070–1079.

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*, 736–753. doi:10.1177/0013164498058005002.

*Květon, P., Jelínek, M., Klimusová, H., & Vobořil, D. (2007). Data collection on the internet: Evaluation of web-based questionnaires. *Studia Psychologica, 49*, 81–88.

*Laranjeira, C. A. (2009). Preliminary validation study of the Portuguese version of the Satisfaction with Life Scale. *Psychology, Health & Medicine, 14*, 220–226.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research and Evaluation, 12*, 1–11.

*Lewis, C. A., Shevlin, M. E., Bunting, B. P., & Joseph, S. (1995). Confirmatory factor analysis of the Satisfaction with Life Scale: Replication and methodological refinement. *Perceptual and Motor Skills, 80*, 304–306.

*Lewis, C. A., Shevlin, M. E., Smékal, V., & Dorahy, M. J. (1999). Factor structure and reliability of a Czech translation of the Satisfaction with Life Scale among Czech university students. *Studia Psychologica, 41*, 239–244.

Manten, A. A. (1973). Scientific literature reviews. *Scholarly Publishing, 5*, 75–89.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37*, 113–115. doi:10.1037/0022-0167.37.1.113.

Mulrow, C. D. (1994). Systematic reviews: Rationale for systematic reviews. *British Medical Journal, 309*, 597–599. doi:10.1136/bmj.309.6954.597.

*Navrátil, M., & Lewis, C. A. (2006). Temporal stability of the Czech translation of the Satisfaction with Life Scale: Test-retest data over one week. *Psychological Reports, 98*, 918–920.

Netemeyer, R., Bearden, W., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. London: Sage.

*Neto, F. (1993). The Satisfaction with Life Scale: Psychometrics properties in an adolescent sample. *Journal of Youth and Adolescence, 22*, 125–134.

Olkin, I. (1996). Meta-analysis: Current issues in research synthesis. *Statistics in Medicine, 15*, 1253–1257.

*Pavot, W., Diener, E., Colvin, C. R., & Sandvik, E. (1991). Further validation of the Satisfaction with Life Scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment, 57*, 149–161.

*Pons, D., Atienza, F. L., Balaguer, I., & García-Merita, M. (2000). Satisfaction with Life Scale: Analysis of factorial invariance for adolescents and elderly persons. *Perceptual and Motor Skills, 91*, 62–68.

Qualls, A. L., & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. *Educational and Psychological Measurement, 56*, 209–214. doi:10.1177/0013164496056002002.

* Sancho, P., Galiana, L., Gutierrez, M., Francisco, E.-H., & Tomás, J. M. (2012). Validating the Portuguese version of the Satisfaction with Life Scale in an elderly sample. *Social Indicators Research* [online first]. doi:10.1007/s11205-012-9994-y

*Shevlin, M. E., & Bunting, B. P. (1994). Confirmatory factor analysis of the Satisfaction with Life Scale. *Perceptual and Motor Skills, 79*, 1316–1318.

*Shevlin, M., Brunsden, V., & Miles, J. N. V. (1998). Satisfaction with Life Scale: Analysis of factorial invariance, mean structures and reliability. *Personality and Individual Differences, 25*, 911–916.

*Siedlecki, K. L., Tucker-Drob, E. M., Oishi, S., & Salthouse, T. A. (2008). Life satisfaction across adulthood: Different determinants at different ages? *The Journal of Positive Psychology, 3*, 153–164.

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., & Maraun, M. D. (2009). Psychometric assessment and reporting practices: Incongruence between theory and practice. *Journal of Psychoeducational Assessment, 27*, 465–476. doi:10.1177/0734282909335781.

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., Ferguson, L. P., Knudsen, J. R. S., & Legere, J. C. (2010). A review of psychometric assessment and reporting practices: An examination of measurement-oriented versus non-measurement-oriented domains. *Canadian Journal of School Psychology, 25*, 246–259. doi:10.1177/0829573510375549.

*Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of quality of life scales: An illustration with the Satisfaction with Life Scale (SWLS). *Social Indicators Research, 92*, 489–496.

*Swami, V., & Chamorro-Premuzic, T. (2009). Psychometric evaluation of the Malay Satisfaction with Life Scale. *Social Indicators Research, 92*, 25–33.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Allyn & Bacon.

The Cochrane Collaboration. (2002). *An introduction to meta-analysis*. Retrieved July 17, 2012, from http://www.cochrane-net.org/openlearning/html/mod3.htm

*Tucker, K. L., Ozer, D. J., Lyubomirsky, S., & Boehm, J. K. (2006). Testing for measurement invariance in the Satisfaction with Life Scale: A comparison of Russians and North Americans. *Social Indicators Research, 78*, 341–360.

*Vautier, S., Mullet, E., & Jmel, S. (2004). Assessing the structural robustness of self-rated satisfaction with life: A SEM analysis. *Social Indicators Research, 68*, 235–249.

*Vittersø, J., Biswas-Diener, R., & Diener, E. (2005). The divergent meanings of life satisfaction: Item response modeling of the Satisfaction with Life Scale in Greenland and Norway. *Social Indicators Research, 74*, 327–348.

*Westaway, M. S., Maritz, C., & Golele, N. J. (2003). Empirical testing of the Satisfaction with Life Scale: A South African pilot study. *Psychological Reports, 92*, 551–554.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58*, 21–37. doi:10.1177/0013164498058001003.

*Wu, C.-H., & Wu, C.-Y. (2008). Life satisfaction in persons with schizophrenia living in the community. *Social Indicators Research, 85*, 447–460. doi:10.1007/s11205-007-9136-0.

*Wu, C.-H., & Yao, G. (2006). Analysis of factorial invariance across gender in the Taiwan version of the Satisfaction with Life Scale. *Personality and Individual Differences, 40*, 1259–1268.

*Wu, C.-H., Chen, L. H., & Tsai, Y.-M. (2009). Longitudinal invariance analysis of the Satisfaction with Life Scale. *Personality and Individual Differences, 46*, 396–401. doi:10.1016/j.paid.2008.11.002.

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. (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: Information Age Publishing.

Chapter 5
Validation Practices in Counseling: Major Journals, Mattering Instruments, and the Kuder Occupational Interest Survey (KOIS)

Eric K.H. Chan, David W. Munro, Alexander H.S. Huang, Bruno D. Zumbo, Roya Vojdanijahromi, and Neelam Ark

Validity theory and validation practice have become more complex during the past half century. Prior to the 1950s, a diversity of procedures was used to study validity and an array of names was used when researchers reported validity evidence; however, the criterion- and content-based models dominated the practice of validation (Anastasi 1986). In 1954, the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* 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.

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 most current edition of the *Standards for Educational and Psychological Testing* (AERA et al. 1999). According to the *Standards* (AERA et al. 1999), validity is "the degree to which

E.K.H. Chan, Ph.D. (✉) • B.D. Zumbo, Ph.D. • N. Ark
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

D.W. Munro • A.H.S. Huang • R. Vojdanijahromi
Counseling Psychology Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada

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_5,

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. The definition of validity, as stated in the *Standards*, espouses construct validity as the central component, and encompasses the following five sources of evidence germane to the investigation of the interpretation and use of the scores of a measure: test content, internal structure, associations with other variables, response processes, and consequences (AERA et al. 1999).

Although these contemporary views have been in the literature for about 25 years, several scholars have questioned whether such contemporary perspectives are reflected in current practices (e.g., Messick 1989; Zumbo 2007). With an aim towards informing validation practices in counseling, the objective of this book chapter was to present the results of three studies synthesizing the practices of validation in counseling in contrast to the modern view of validity as endorsed by the *Standards* (AERA et al. 1999). The three studies are meant to provide a broad view of validation practices in counseling. In Study 1, validation papers published in four major counseling journals (*Journal of Counseling Psychology, The Counseling Psychologist*, *Journal of Counseling and Development*, and the *Canadian Journal of Counseling and Psychotherapy*) in North America in 2009, 2010, and 2011 were synthesized. The next two studies investigated validation practices for two constructs in counseling: one relatively new (i.e., mattering) and the other a measure with a long tradition in counseling (career assessment). In Study 2, the validation practices of papers of the mattering construct were summarized. In Study 3, the validity evidence of the Kuder Occupational Interest Survey (KOIS) was reviewed.

Examining the reporting characteristics of validity articles published in different areas in counseling is a useful method to study the practices of validation in this academic discipline. It is important to note that our goal was not to evaluate the quality of the existing instruments in counseling. Instead, our focus was on mapping out the current available validation practices with an aim to informing future validation practices in counseling.

Study 1 – Four Counseling Journals

Counseling psychologists and counselors often use psychometric instruments for assessing career interests, mental health functioning, and the effectiveness of counseling interventions (Nugent and Jones 2005). Measurement validity is an important area in counseling (Bolt and Rounds 2000) and is a fundamental issue in evaluating the psychometric properties of instruments (AERA et al. 1999).

A large number of measurement validity studies have been published in the counseling literature. However, in a review of papers reporting the measurement properties of measures published in the *Journal of Counseling Psychology* in 1967, 1977, and 1987, Meier and Davis (1990) found that the evidence to support the

validity of the instruments used was inadequately reported. The purpose of Study 1, therefore, was to investigate the reporting practices of validity evidence in papers published in recent issues in prominent counseling journals in North America, and to assess the extent to which the practice of validation is consistent with the modern thinking in validity theory as reflected in the most current issue of the *Test Standards* (AERA et al. 1999).

Method

Validity papers published in four major counseling journals in North America in 2009, 2010, and 2011 were reviewed. The four journals included:

1. *Journal of Counseling Psychology* (an American Psychological Association [APA] journal);
2. *The Counseling Psychologists* (official journal of Division 17 [counseling psychology] of the APA);
3. *Canadian Journal of Counseling and Psychotherapy* (official journal of the Canadian Counseling Association); and
4. *Journal of Counseling and Development* (official journal of the American Counseling Association).

A systematic search was conducted using the official website of each of the four journals. Articles with the keywords "valid", "validity" or "validation" in the title or abstract were retrieved and reviewed in detail. To be included in the present analysis, the study must explicitly state that validity is the focus/objective and be empirical studies. We excluded opinion papers, editorials, reviews, systematic reviews, meta-analyses, as well as guidelines, recommendations, and expository papers about statistical methods.

Building on previous research by Cizek and colleagues (2008, 2010), a coding form was developed with the following sources of validity evidence: face, content, construct, predictive, concurrent, convergent, discriminant, response processes, and consequences. These reflect the five major sources of validity evidence in the framework proposed by Messick (1989), which is also stated in the most current edition of the *Test Standards* (AERA et al. 1999). The five sources include (a) content-related, (b) response processes, (c) internal structure, (d) associations with other variables, and (e) consequences. We also coded reliability evidence, including internal consistency, test-retest, and inter-rater.

The coding for Study 1 was based on the sources of validity evidence that the authors reported as well as the validation methodology employed. For instance, if "discriminant validity" was explicitly stated in an article, discriminant validity was coded. If factor analytic results were reported but were not explicitly referred to as internal structure evidence in the paper, we coded the evidence as internal structure.

Results and Discussion

A total of 21 papers met our inclusion criteria. The number of sources of validity evidence reported per study ranged from zero to five, with a mode of two. About 95 % of the studies reported between two to five sources of validity evidence (see Table 5.1). Table 5.2 presents the sources of validity evidence reported in the four journals. Given that many of the papers reported more than one source of validity, the sum of the percentages across sources will not equal 100 %. Internal structure was the most frequently reported source of validity evidence followed by construct validity, convergent validity and discriminant validity. Content validity and concurrent validity were each reported in approximately one third of the studies. Face validity, although not regarded as a source of validity, was presented in one study. Predictive validity, response processes, and consequences were not reported in any of the articles. No study approached validation from the unitary perspective.

With respect to the reported reliability evidence, approximately 90 % of the 21 articles reported evidence on internal consistency. Slightly less than two thirds

Table 5.1 Frequency of number of validity sources reported in four major counseling journals

Number of sources	Frequency	Percent
0	1	4.80
1	0	0
2	5	23.80
3	5	23.80
4	6	28.60
5	4	19.00
Total	21	100

Table 5.2 Sources of validity evidence reported in four major counseling journals

Validity	Number	Percent
Internal structure	20	95.24
Construct	14	66.67
Convergent	13	61.90
Discriminant	9	42.86
Content	6	28.57
Concurrent	6	28.57
Face	1	4.76
Predictive	0	0
Response processes	0	0
Consequences	0	0
Reliability	Number	Percent
Internal consistency	19	90.48
Test-retest	13	61.90
Inter-rater	0	0

A paper can report more than one source of validity

of the articles reported test-retest reliability evidence, while no article reported evidence on inter-rater reliability.

The results of the present study showed that a broad perspective on the possible sources of validity evidence is reported in the articles published in the four major counseling journals in North America. Researchers conducting validation studies (in this instance, within the realm of counseling) are not relying on only one source of validity evidence to the exclusion of all other sources; however, some sources of validity evidence, such as response processes and consequences, are absent. We also observed that the modern view of validity (as stated in the *Test Standards*) is not having a strong presence across the major journals in counseling in North America. This is particularly noteworthy given that two journals are APA journals and the APA has been a driving force for the *Standards* since there earliest inception in the 1950s. There seems to be an important disconnect, therefore, between validity theorists and validation practitioners.

Study 2 – Mattering Instruments

Each individual has an inherent need to feel that they are noticed by, important to, and cared for by others – that they *matter* (Rosenberg and McCullough 1981). Mattering was conceptualized by Rosenberg and McCullough (1981) as "a motive: the feeling that others depend on us, are interested in us, are concerned with our fate, or experience us as an ego extension" (p. 165). Mattering has more recently been defined as "the perception that, to some degree and in any of a variety of ways, we are a significant part of the world around us" (Elliot et al. 2004, p. 339). The opposite of mattering is a sense of insignificance and irrelevance in a hostile world (Elliot et al. 2004). When individuals perceive that they do *not* matter, thus feeling irrelevant and uncared for, it is difficult to develop a healthy self-concept (Rosenberg and McCullough 1981), and it may increase the likelihood that they engage in maladaptive behaviors in order to seek a sense of significance in the world (Elliot et al. 2004). William James (1890) noted the importance of mattering over 100 years ago when he stated the following: "No more fiendish punishment could be devised, were such a thing physically possible, than that one should be turned loose in society and remain absolutely unnoticed" (p. 293).

Mattering begins interpersonally but registers intrapersonally, ultimately affecting one's self-concept. Awareness that one matters to others boosts feelings of relatedness and one's sense of meaning and purpose in life (Marshall 2001). However, in order to be effective, the receiver of mattering behavior has to subjectively perceive that he or she matters. If one had experienced "not-mattering" in the past, it could be hypothesized that he or she would be less likely to perceive, even though objectively evident, that someone is behaving as though he or she matters (Elliot et al. 2004).

From the time that Rosenberg (1989) first conceptualized the term until the beginning of the new millennium, there has been little research on mattering. Since

2001, however, mattering has been studied with a wide range of populations: adolescents, college students, adults, couples, employees with and without mental illness, medical residents, military cadets, and school counselors (Connolly and Myers 2002; Corbiere and Amundson 2007; Dixon et al. 2009; Dixon Rayle 2005; Elliott et al. 2004, 2005; France and Finney 2009; Mak and Marshall 2004; Marshall 2001; Myers and Bechtel 2004; Powers et al. 2004). Research indicates that mattering to others is positively correlated with self-esteem and well-being, and negatively correlated with anxiety, depression, academic difficulties, suicidal ideation, hostility, and aggression (Dixon Rayle 2005; Elliot et al. 2004; Elliot 2009; Marshall 2001).

The development of scales to measure mattering has progressed along several different pathways, including relationship counseling, career counseling, school counseling, and health care providers. The purpose of the present study was to review the validity evidence reported for mattering measures and compare the validation practices in mattering against contemporary views of validity (e.g., Kane 2006; Messick 1989) and as reflected in the *Standards for Educational and Psychological Testing*, hereafter referred to as the *Standards* (AERA et al. 1999).

Measures of Mattering

Whereas most of the studies on mattering cite Rosenberg and McCullough's (1981) initial work, some rely more heavily on the investigations of other early mattering scholars from the late 1980s and early 1990s to inform their scale construction. The measures developed are unidimensional or multidimensional, and all of the studies except one (Corbiere and Amundson 2007) utilized samples of adolescent or young adult students in Canada or the United States.

Marshall (2001) devised the unidimensional Mattering to Others Questionnaire (MTOQ), consisting of 11 items assessing adolescents' perceptions of mattering to parents and friends. Later, she and a colleague adapted the MTOQ in order to create the Mattering to Romantic Others Questionnaire, a 17-item scale that measures adolescents' perceptions of mattering to romantic partners (Mak and Marshall 2004).

Within the field of career counseling, Corbiere and Amundson (2007) investigated the applicability of the multidimensional Ways of Mattering Questionnaire, developed by Amundson (1993), for individuals with mental illness who were participating in supported employment programs.

Elliot et al. (2004) constructed a three-factor 24-item index of mattering, while France and Finney (2009) further refined Elliot et al.'s (2004) index from three to four factors and later devised the University Mattering Scale (France and Finney 2010). Tovar et al. (2009) also constructed a scale to measure mattering in postsecondary institutions. However, in addition to accounting the work of Rosenberg and McCullough (1981), Tovar et al. (2009) cite the Mattering Scales for Adult Students in Postsecondary Education (Schlossberg 1990) as being a

model for their College Mattering Inventory, a six-factor scale consisting of 29 items. In contrast, Guirguis and Chewning (2008) utilized the General Mattering Scale (Marcus 1991) and incorporated an adapted version of this scale within a measure to assess pharmacy students' beliefs regarding the monitoring of chronic diseases. All in all, mattering is an important concept in counseling research and practice.

Method

The present study was focused on papers that provided empirical evidence of the validity of mattering measures and hence explicitly positioned themselves as validation research – as opposed to papers that incidentally presented what might be considered, more generally, validity evidence. Articles for review were identified through a comprehensive search of the ERIC and PsycINFO databases conducted in August 2011 using "mattering" AND "validity" as keywords. The search was not restricted by year of publication. Research articles were included in the present review if they provided empirical information on the validity evidence of a measure of mattering and were published in peer-reviewed journals. Building on previous similar research (Auewarakul et al. 2005; Beckman et al. 2005; Chan et al. 2011; Chan and Zumbo 2012; Cizek et al. 2008, 2010), the five sources of validity evidence as stated in the *Standards* (AERA et al. 1999) were coded and included (a) content-related, (b) response processes, (c) internal structure, (d) associations with other variables, and (e) consequences.

Results and Discussion

Eight studies were included in the present review. Although the search was not restricted by year of publication, all of the studies identified for inclusion were published in the past decade (2001–2011). Thus, although the concept of mattering was first devised in the 1980s (Rosenberg and McCullough 1981), there has been scant attention paid to the validity of measures of mattering until the past decade.

None of the studies included in this review cited a modern framework for validity evidence, such as the one proposed by Messick (1989) and espoused in the *Standards* (AERA et al. 1999). Although Marshall (2001) cited Loevinger's (1957) three stages for obtaining validity evidence (substantive, structural, and external) and stated that Messick (1989) outlined a similar means of finding validity evidence, Marshall (2001) did not incorporate Messick's (1989) vocabulary. Only one study (France and Finney 2009) referred to obtaining validity evidence for the interpretation and use of scores, the importance of which are articulated in the *Standards* (AERA et al. 1999). The remaining articles stated that they were establishing the validity of a measure, which is indicative of adherence to the

traditional view of validity and therefore not in line with the contemporary unitary paradigm of validity.

The most frequently reported sources of psychometric evaluation were internal structure (all eight studies) and internal consistency reliability (six studies). Two out of eight studies explicitly mentioned the examination of convergent validity (Corbiere and Amundson 2007; Tovar et al. 2009). Five other studies also looked at convergent validity but did not label it as such (Elliot et al. 2004; France and Finney 2009, 2010; Mak and Marshall 2004; Marshall 2001). One study reported investigating content validity (Elliot et al. 2004), while another study described the examination of content validity without actually providing the term in its report (Marshall 2001). Finally, only one study investigated face validity (Tovar et al. 2009), and one study reported evidence on discriminant validity but did not label it as such. None of the studies explicitly mentioned predictive validity, concurrent validity, response processes, consequences, or cited a contemporary validity reference such as the *Standards* (AERA et al. 1999).

The present review showed that validity information in mattering measures was not reported using the contemporary view of validity based on Messick's (1989) work as put forward by the *Standards* (AERA et al. 1999). These findings are consistent with those of Cizek and colleagues (2008, 2010) who reviewed the validity evidence in papers published in the *Mental Measurement Yearbook*, education journals, and presentations in major education and psychology conferences. Response processes and consequences of test use were virtually non-existent in the articles reviewed in the present study. A closer look at the reviewed articles with specific regard to the five sources of validity evidence as stated in the *Standards* is discussed in the following section with an aim toward informing the validation practices of existing mattering measures and the measures thereof.

Messick (1989) reminded us that 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). Validity evidence from a number of different sources (i.e., the five sources) is needed to support the interpretation of scores. Although Messick's (1989) view has been around for over 20 years, the results of our review showed that his view of validity has not permeated the validation practices in mattering measurement.

Similar to the previously presented study on the systematic review of four major counseling journals, the central message of this study is that there is a discrepancy between validity theory and validation practices. In this instance, the area concerns the construct of mattering, and there is a divergence between what validation practitioners report and what Messick (1989) and the *Standards* (1999) have put forward as general guidelines regarding the practice of validation. Within this review, internal structure was the most widely reported source of validity evidence; however, certain sources of validity evidence, such as response process and consequences, were not being reported in measures of mattering. Validation methods have been discussed and applied by various authors in these under-studied sources of validity evidence (e.g., Barofsky et al. 2003; Gadermann et al. 2011; Hubley and Zumbo 2011; Reeve et al. 2011; Zumbo 2007, 2009). More validity evidence,

particularly evidence on response processes and consequences, are needed to strengthen the interpretation and inferences made from the scores from mattering measures.

Mattering is instrumental in the development of one's self-concept and sense of belonging (Elliot et al. 2004), which are both integral components of the self, whose growth may be stymied by non-mattering experiences. Measures with strong validity evidence to measure this important sociological and psychological concept should be based on the most up-to-date methods of conducting research and also adhere to a standardized and modernized means of reporting results so that research in this relatively new and promising area can grow coherently and consistently. More work is needed to strengthen the validity of score interpretations in the measurement of mattering. Perhaps more communication between validity theorists and practitioners who do validation work in the area of mattering is needed.

Study 3 – The Kuder Occupational Interest Survey

Within career counseling, a widely used vocational measure, developed by Frederic Kuder in 1966 (Zytowski 1992), is the Kuder Occupational Interest Survey (KOIS). The measure evolved from Kuder's work on interests during the depression era; thus, the history of the scale dates back nearly 80 years (Diamond and Zytowski 2000). The self-report measure aims to determine the shared interests of the respondent with two criterion group scales: an occupational scale and a college major scale. The criterion groups are created through a technique devised by Strong (1943) for the Strong Vocational Interest Blank (SVIB). More specifically, the criterion group scales are derived from the "satisfaction" claimed by individuals who possess a level of experience (indicated by number of years) within a particular occupation or college major (Diamond and Zytowski 2000). These individuals state they would select the same profession or college major again if they were to repeat the selection process (Zytowski 1992). As a result, their specific interests are determined, and the respondent's scores are compared to those within the criterion group scales (Zytowski 1992). Interests are discerned through 100 forced-choice triad activity questions, "each requiring respondents to indicate which of three activities they prefer most and which they prefer least" (Diamond and Zytowski 2000, p. 263).

The intended population of the measure is for high-school age and older individuals, and the scale is often administered within vocational counseling settings (Zytowski 1992). For the respondent, a final report is generated that is comprised of four sections: Dependability, Vocational Interest Estimates (VIEs), Occupations, and College Majors (Zytowski 1992). Dependability refers to the degree of confidence that can be placed on the respondent's ability to answer items in a sincere and accurate manner (Zytowski 1992). This section aims to ascertain whether the respondent was disingenuous, such as answering the items by chance or for social desirability reasons. Dependability is determined primarily through a "Verification

(V) scale," a measure within the KOIS that involves 74 response positions that have "low frequencies of most or least endorsements. Thus, someone who answered carelessly or by chance would [likely] endorse more of the items in the V scale" (Zytowski 1992, p. 246). The second section of the report, VIEs, revolves around ten core interests: outdoor, clerical, mechanical, musical, social service, literary, computational, persuasive, scientific, and artistic (Diamond and Zytowski 2000). The respondent receives two sets of VIE scores: "one based on male norms and one based on female norms" (Zytowski 1992, p. 247). The final two sections of the report involve a comparison of the respondent's scores with the scores of men and women within particular occupations and college majors. These sections highlight how closely the respondent's scores resemble men and women within the two criterion groups (Diamond and Zytowski 2000). The justification for comparing the respondent's scores to men and women, in both the VIEs and the two criterion group scales, is that sex-role socialization and distinct expectations between the genders have resulted in different responses to the activities characterized by the test items (Diamond and Zytowski 2000).

The interpretations and judgments generated from the KOIS final report can have a significant impact on a client's life circumstances. For instance, while the purpose of the KOIS is to offer the respondent an interest profile and comparative data, an individual may actually pursue a particular occupation because his/her interests are highly similar to a specific profession. As the interpretation of KOIS scores has potential consequences on people's career choice, Study 3's objective, like Studies 1 and 2, was to examine the validity evidence reported for the KOIS and to compare the validation practices against contemporary views of validity.

Method

We searched for articles using the following five databases: ERIC, PsycINFO, PsycARTICLES, PsycEXTRA, and Academic Search Complete. The following keywords were used: "Kuder Occupational Interest Survey", "KOIS", "validity", "validation", "psychometric", "evidence", "reliability", and "measurement". With "Kuder Occupational Interest Survey" and "KOIS" as the keywords within the first search field, and the remaining keywords within the second search field used with the "Title-field" option, a total of sixteen articles were retrieved. These articles span four decades, 1970–2010.

To decide on which articles to select, several inclusion/exclusion criteria were used. First, articles had to either mention validity in the title, or position themselves as validity studies in the abstract. In addition, we focused only on peer-reviewed articles, and only English-language papers were selected. Duplicates, reviews of the KOIS, dissertations, articles that assessed the validity of other vocational scales, and articles that discussed aspects outside the scope of validity evidence were excluded. Consequently, six articles were excluded from the final analysis, and the ten remaining articles were all chosen for analysis in the systematic review.

The ten articles were coded according to a coding scheme developed by the authors. The coding scheme focused on the five sources of validity evidence proposed by the *Standards* (AERA et al. 1999). This coding scheme, therefore, included the following categories: (a) content validity, (b) association with other variables (predictive, concurrent, convergent, discriminant validity, and construct validity), (c) consequences, (d) internal structure, and (e) response processes. In addition, an "other validity" category was included to account for validity sources not associated with the evidence sources described by the *Standards* (AERA et al. 1999). Finally, four reliability categories (internal consistency, test-retest, inter-rater, and parallel forms) and an "other reliability" category were included. In total, 15 headings were created in relation to sources of validity evidence.

Results and Discussion

Table 5.3 summarizes our findings by presenting the articles, and the percentages, that reported on a specific source of validity evidence. Of the ten articles included in the present review, five articles reported evidence on predictive validity, while three articles addressed concurrent validity. Three separate articles addressed convergent, discriminant, and construct validity, indicating a representation of 10 %, respectively, for each source of evidence. [Note that construct validity was linked to association with other variables because the definition ascribed to this source of validity evidence was not consistent with the one outlined by Messick (1989) or the *Standards* (AERA et al. 1999)]. One article reported on internal structure, while content validity, response processes, and consequences were not reported in any of the articles.

Analysis of the articles also identified three sources of validity evidence outside the scope advocated by Messick (1989) or the *Standards* (AERA et al. 1999); specifically, one article reported on face validity, one for congruent validity, and one for accuracy-as-classification. "Face validity" was reported through an interpretation of "reasonableness," or whether the participants agreed with the results indicated by their final report (Denker and Tittle 1976). The process of assessing "congruent validity" involved comparing two vocational measures, the KOIS and the SVIB, and determining whether equivalently named interest scales within each measure, such as "Lawyer" or "Accountant," attained similar scores for respondents (Zytowski 1972a, b). Finally, "accuracy-as-classification" was described as a source of validity evidence designed specifically for vocational measures. It was used to determine whether the measure would suggest the same occupation for individuals who are working in established vocations (Zytowski 1972a, b). For example, this source of validity evidence aims to determine whether the KOIS is able to suggest "lawyer" as a viable profession in the final report for a practicing lawyer.

Finally, with respect to the reported reliability evidence, one article reported on test-retest and one article addressed internal consistency. Additionally, there were

Table 5.3 Sources of reported validity evidence for the KOIS

Validity	Number	Percent
Content validity	0	0
Association with other variables		
(a) Convergent	1	10.00
(b) Discriminant	1	10.00
(c) Construct	1	10.00
(d) Criterion-related		
Predictive	5	50.00
Concurrent	3	30.00
Internal structure	1	10.00
Response processes	0	0
Consequences	0	0
Other validity		
Face validity	1	10.00
Congruent validity	1	10.00
Accuracy-as-classification	1	10.00
Reliability	**Number**	**Percent**
Test-retest	1	10.00
Internal consistency	1	10.00
Inter-rater reliability	0	0
Parallel forms	0	0
Other reliability		
Absolute	1	10.00
Intra-individual	1	10.00

two other reliability sources addressed: one article on absolute and one article on intra-individual. Both absolute and intra-individual refer to the stability of a psychological construct over time (Rottinghaus et al. 2007). In particular, Rottinghaus et al. (2007) first examined the stability of interests over time at an "absolute" level by comparing the mean scores on the KOIS final report for subjects during two time intervals. The researchers then assessed "intra-individual" stability by examining whether a particular respondent's interest profile remained consistent over time (Rottinghaus et al. 2007).

In order to address the question of whether the available validity evidence resembles the modern view of validity, two types of analyses were conducted. The first involved uncovering the year the articles were published. In particular, a comparison was made between the number of articles published prior to Messick (1989) and the *Standards* (AERA et al. 1999). This aspect was seen as integral because information from these two sources may not have been available for the researchers during their analysis of validity evidence for the KOIS; consequently, the contemporary perspective of validity may not have been emphasized.

Analysis indicated that the majority of validity evidence articles (seven) were assessed prior to either Messick (1989) or the *Standards* (AERA

et al. 1999). In contrast, only three articles were found to contain sources of validity evidence for the KOIS after Messick (1989) and the *Standards* (AERA et al. 1999). It appears that assessing the validity evidence of the KOIS was prolific during the period prior to the contemporary perspective of validity, from 1972 to 1979. Afterwards, the studies became scarce, and only three articles were found from 1998 to 2007. Information on the contemporary perspective of validity, therefore, was presumably unavailable for the majority of articles within this review that assessed the validity evidence of the KOIS. This finding supports the observations made by Savickas et al. (2002) who emphasize that research into the validity evidence of the KOIS was "productive during the 1960s and 1970s," yet "lay mostly dormant during the 1980s and 1990s" (p. 140).

The second analysis conducted was to determine the validity perspective subscribed to by the articles. This method of analysis followed the criteria employed by Cizek et al. (2008) who utilized three indicators for an article's validity perspective: (1) whether articles mentioned a unitary perspective of validity, (2) whether articles cited Messick (1989) or the *Standards* (AERA et al. 1999), and (3) whether articles stated that validity was a property of the test or a property of test scores. For determining whether a unitary perspective of validity was supported within a particular article, a similar method to the one devised by Cizek et al. (2008) was utilized. It was assumed that if a particular publication advocated a unified view of validity that emphasizes all sources of validity evidence as construct validity, then it would have mentioned this perspective in the article (Cizek et al. 2008). The second mandate for citing Messick (1989) or the *Standards* (AERA et al. 1999) was restricted to three articles because only three articles were published after 1989. The third mandate was assessed through the language employed within the articles.

Analysis of validity perspectives indicated that none of the articles appeared to reflect the unitary perspective. Indeed, as one author notes, while any measure "must be shown to possess validity, [vocational measures] need to demonstrate the predictive kind" (Zytowski 1976, p. 221). The use of the word "kind" illustrates that some of the analyzed articles appeared to view validity as composed of several types, rather than as a unified concept. While this finding is not surprising for the seven articles published prior to Messick (1989) or the *Standards* (AERA et al. 1999), it is worth noting that the contemporary articles also do not appear to subscribe to a unified perspective of validity. This may explain why none of the three articles published after 1989 made reference to either Messick (1989) or the *Standards* (AERA et al. 1999).

When a concept of validity was emphasized, seven of the articles conceived of validity as an aspect of the measure. Indeed, phrases used to describe the source of validity evidence were often attributed to the KOIS itself rather than referencing judgments or inferences made from the scores on the final report (e.g., Zytowski 1972a, b, 1976). In contrast, only one article made reference to validity as an aspect of scores or interpretations. The article specifically states that predictive validity is determined by assessing whether "interest scores match one's future occupation" (Rottinghaus et al. 2007, p. 7). The remaining two articles were either unclear or did not mention a concept of validity. These findings are not surprising given that most

of the sources of validity evidence in this review were published prior to Messick (1989) or the *Standards* (AERA et al. 1999). As a result, they may have been more inclined to endorse a view of validity more in agreement with a previous conception that highlighted validity as a characteristic of a measure, rather than as a characteristic of inferences or interpretations derived from the measure's scores.

Our findings showed that there seems to be discrepancies between the validity evidence for the KOIS and the *Standards* (AERA et al. 1999). Most articles did not mention the unitary perspective of validity and none cited Messick (1989) or the *Standards* (AERA et al. 1999). In addition, the characterization of validity was often seen as a part of the measure rather than in relation to score interpretations.

Discussion

Our findings indicate that validity evidence available for the KOIS predominantly focuses on association with other variables. There is a strong emphasis placed on predictive validity and concurrent validity evidence. Additionally, while one article addressed internal structure, the selected articles did not attend to any of the remaining sources of validity evidence for demonstrating construct validity (response processes, consequences, and test content were all not present). When construct validity was referenced, the definition was distinct from the one outlined by Messick (1989) or the *Standards* (AERA et al. 1999). For example, construct validity was seen as determined by discriminant and convergent validity evidence (Savickas et al. 2002). This particular definition, therefore, neglects the majority of sources of validity evidence, and in fact, only addresses two features of associations with other variables. There was also a focus on three validity sources unassociated with the evidence for construct validity: face validity, congruent validity, and accuracy-as-classification. The reliability evidence available for the measure was also quite limited, with 10 % reported on test-retest reliability and internal consistency respectively. Finally, one article addressed two other reliability sources, labelled as absolute and intra-individual stability.

The perspective of validity reflected by the majority of the articles also appears to be in disagreement with the contemporary perspective of validity. The important validity evidence described by many of the articles stress predictive validity, instead of construct validity. Furthermore, the unitary view of validity was not endorsed in any of the articles, as many reported on "kinds" of validity. The concept of validity was also often perceived as a characteristic of the KOIS rather than as an aspect of the interpretations generated from the scores, and none of the articles cited Messick (1989) or the *Standards* (AERA et al. 1999). The latter finding was relevant to the three articles published after 1989. Finally, the subjective assessment indicated that that validity evidence reported by the articles may be inaccurately labelled.

In conclusion, much like those conducting validation research within the domain of counseling, it appears that KOIS validation researchers have not fully endorsed a

unitary perspective of validity. The findings of the present review have the potential to stimulate more validation research in this area.

General Discussion

In this book chapter, we presented the results of three studies synthesizing the validation practices in the area of counseling research and examined the extent to which the reported validity evidence aligns with Messick's (1989) modern view of validity as endorsed by the *Test Standards* (AERA et al. 1999).

The results of the present study revealed that a broad perspective of measurement validity is represented in the validation papers published in counseling journals, indicating that counseling researchers conducting validation studies are not relying on only one source of validity evidence at the exclusion of all others. However, some sources of validity evidence such as response processes and consequences are absent, and the modern view of validity does not have a strong presence in the validation practice within counseling.

Response processes refer to the thinking or cognitive processes involved when a client or research participant responds to items on a measure (AERA et al. 1999; Messick 1989, 1995). In other words, the purpose is to investigate how and why people respond to questions or items the way they do. Consequences were also not reported in the counseling literature we reviewed. Consequences in validity refer to (1) the intended use of measure scores and (2) the misuse of measure scores (AERA et al. 1999; Hubley and Zumbo 2011, 2013; Messick 1989). Both response processes and consequences are emerging as central to claims of measurement validity (Hubley and Zumbo 2013; Messick 1995) and is important in achieving a strong form of construct validity by understanding how theory helps *explain* the variation in test scores (Zumbo 2007, 2009). Consequences are particularly relevant in counseling research because of the scope of practice and the connection to the ethics of assessment. In addition, a consideration of consequences as described in Hubley and Zumbo (2011, 2013) sets the stage for a framing of validation practices in a context wherein value implications, intended social consequences, and unintended side effects of legitimate test interpretation and use are highlighted – all of which are key issues in assessment in counseling.

Our findings also reveal the lack of qualitative studies in validation work in counseling. For instance, in the examination of response processes, qualitative methods such as the "think aloud" procedure are used in other areas of research to establish what Messick referred to as the 'substantive' aspect of construct validity (e.g., Gaderman et al. 2011). Counseling researchers are well versed in qualitative research methods and such methods are important in measurement validation research. Validation research in counseling should include both qualitative and quantitative methods.

Our results from the three studies presented in this book chapter suggest a discrepancy between validation practices and contemporary validity theory. There

exists a stark contrast between validity evidence presented and the contemporary theoretical perspective advocated by the *Test Standards* (AERA et al. 1999). We provide a few possible reasons for such a discrepancy.

First, researchers and practitioners may neglect the predominant theoretical orientation because they are unaware of the contemporary unitary view of validity. Second, there may be an inclination towards determining a specific source of validity evidence because it is seen as the primary purpose of the measure itself. For example, if the measure is a predictive tool, such as the KOIS, prediction studies are seen as the key evidence. One intended consequence is for the respondent of the KOIS to explore certain occupations and college majors, and perhaps even pursue them in the future. There remains, therefore, a pressure for the KOIS to accurately reflect the matched interests of the respondent to particular occupations and college majors; otherwise, the utility of the measure may be seen as fruitless. It is not surprising, then, that the majority of the reviewed articles addressed predictive validity in some manner, to the neglect of others.

In addition, some sources of validity evidence may be difficult to attain; it may be difficult to acquire data for response processes for measures, especially when a scale possesses numerous items. For instance, the KOIS contains 100 items; therefore, it may be perceived as cumbersome and time-consuming for practitioners to assess the individual thoughts that underlie a respondent's decision-making process for choosing the most preferred activity, the least preferred activity, and the reason one activity was disregarded. As a result, this aspect may also be ignored.

Future Directions and Recommendations

A few ways to improve the practice of validation and reporting of validity evidence in counseling are recommended. First, improving the transparency, accuracy, and completeness in the reporting of validity studies is necessary. As Meier and Davis (1990) commented, a lack of standards for the reporting of validity evidence may be one reason that explains the suboptimal reporting of such evidence in counseling. We believe guidelines with a set of recommended items that authors should report for their validity studies are needed (see Chap. 5 of this edited book). Having accepted and endorsed reporting guidelines on validity would allow the standardization of information reported in validity studies and improve the quality of the peer review process.

Second, if reporting guidelines for validation studies are established, they need to be adopted by researchers, journal editors, journal reviewers, and the broader academic community. The use of accepted reporting guidelines is associated with better quality academic publications (Cobo et al. 2011). Journal editors play an important role in the peer review process; they are therefore in the best position to promote the use of guidelines for the reporting of validity evidence.

Third, more concerted efforts are needed to expand the graduate curriculum to include courses or seminars in validity theory. In a survey of doctoral training in

psychology, Aiken et al. (2008) found that the median amount of time devoted to training in statistics, measurement, and methodology was 1.6 years, with a mode of 1 year. They also found that the main focus was introductory-level statistics, with little coverage of advanced statistics and measurement (Aiken et al. 2008). Given the relatively limited amount of time devoted to measurement, it is reasonable to believe that graduates may not be aware of the modern unitary view of validity. Graduate programs need to incorporate into their curricula advanced-level measurement courses covering the modern view of validity.

References[1]

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.[*1] [*2] [*3]

Amundson, N. E. (1993). Mattering: A foundation for employment counseling and training. *Journal of Employment Counseling, 30*, 146–152.[*2]

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

Auewarakul, C., Downing, S. M., Jaturatamrong, U., & Praditsuwan, R. (2005). Sources of validity evidence for an internal medicine student evaluation system: An evaluative study of assessment methods. *Medical Education, 39*, 276–283.[*2] [*3]

Bailey, T. K. M., Chung, Y. B., Williams, W. S., Singh, A. A., & Terrell, H. K. (2011). Development and validation of the Internalized Racial Oppression Scale for Black individuals. *Journal of Counseling Psychology, 58*(4), 481–493.[*1] [#]

Barofsky, I., Meadows, K., & McColl, E. (2003). Cognitive aspects of survey methodology and quality of life assessment: Summary of meeting. *Quality of Life Research, 12*, 281–282.[*2]

Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching. *Journal of General Internal Medicine, 20*, 1159–1164.[*2] [*3]

Bolt, D. M., & Rounds, J. (2000). Advances in psychometric theory and methods. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (3rd ed., pp. 140–176). Hoboken: Wiley.[*1]

Brewster, M. E., & Moradi, B. (2010). Perceived experiences of anti-bisexual prejudice: Instrument development and evaluation. *Journal of Counseling Psychology, 57*(4), 451–468.[*1] [#]

Chan, E. K. H., & Zumbo, B. D. (2012, April). *When validity theory meets validation practice: Research syntheses of validity evidence reported in seven areas*. Symposium to be conducted at the Annual Convention of the American Educational Research Association, Vancouver, BC, Canada.[*2]

[1] **Note:** The following notations are used to denote specific references

[*1] *indicates references from Study 1; additionally, references marked with* [#] *indicate studies included in the review*

[*2] *indicates references from Study 2; additionally, references marked with a circle* [o] *indicate studies included in the review*

[*3] *indicates references from Study 3; additionally, references marked with a square* [□] *indicate studies included in the review*

Chan, E. K. H., Darmawanti, I., Mulyana, O. P., & Zumbo, B. D. (2011). Validity evidence and validation practice in papers published in *Value in Health* (1998–2010). *Value in Health, 14*, A151.[*2]

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.[*1 *2 *3]

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.[*1 *2 *3]

Cobo, E., Cortés, J., Ribera, J. M., Cardellach, F., Selva-O'Callaghan, A., Kostov, B., et al. (2011). Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: Masked randomized trial. *British Medical Journal, 343*, d6783.

Connolly, K. M., & Myers, J. E. (2002). Wellness and mattering: The role of holistic factors in job satisfaction. *Journal of Employment Counseling, 40*, 152–159.[*2]

Corbiere, M., & Amundson, N. (2007). Perceptions of the ways of mattering by people with mental illness. *The Career Development Quarterly, 56*, 141–149.[*2 o]

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

Del Prado, A. M., & Church, A. T. (2010). Development and validation of the Enculturation Scale for Filipino Americans. *Journal of Counseling Psychology, 57*(4), 469–483.[*1 #]

Denker, E. R., & Tittle, C. K. (1976). " Reasonableness" of KOIS results for re-entry women: Implications for test validity. *Educational and Psychological Measurement, 36*(2), 495–500.[*3 □]

Diamond, E. E., & Zytowski, D. G. (2000). The kuder occupational interest survey. In C. E. Watkins & V. L. Campbell (Eds.), *Testing and assessment in counselling practice* (pp. 263–294). Mahwah: Lawrence Erlbaum Associates.

Dixon Rayle, A. L. (2005). Adolescent gender differences in mattering and wellness. *Journal of Adolescence, 28*, 753–763.[*2]

Dixon, A. L, Scheidegger, C., & McWhirter, J. J. (2009).The adolescent mattering experience: Gender variations in perceived mattering, anxiety, and depression. *Journal of Counseling & Development, 87*, 302–318.[*2]

Einarsdóttir, S., Rounds, J., & Su, R. (2010). Holland in Iceland revisited: An emic approach to evaluating US vocational interest models. *Journal of Counseling Psychology, 57*(3), 361–367.[*1 #]

Elliott, G. C. (2009). *Family matters: The importance of mattering to family in adolescence*. Chichester: Wiley-Blackwell.[*2]

Elliott, G. C, Kao, S., & Grant, A. M. (2004). Mattering: Empirical validation of a social-psychological construct. *Self and Identity, 3*, 339–354.[*2 o]

Elliott, G. C, Colangelo, M. R., & Gelles, R. J. (2005). Mattering and suicide ideation: Establishing and elaborating a relationship. *Social Psychology Quarterly, 68*, 223–238.[*2]

Fan, J., Meng, H., Gao, X., Lopez, F. J., & Liu, C. (2010). Validation of a US Adult Social Self-Efficacy Inventory in Chinese Populations 1Ψ7. *The Counseling Psychologist, 38*(4), 473–496.[*1 #]

France, M., & Finney, S. J. (2009). What matters in the measurement of mattering? A construct validity study. *Measurement and Evaluation in Counseling and Development, 42*, 104–120.[*2 o]

France, M., & Finney, S. J. (2010). Conceptualization and utility of university mattering: A construct validity study. *Measurement and Evaluation in Counseling and Development, 43*, 48–65.[*2 o]

Friedlander, M. L., Friedman, M. L., Miller, M. J., Ellis, M. V., Friedlander, L. K., & Mikhaylov, V. G. (2010). Introducing a brief measure of cultural and religious identification in American Jewish identity. *Journal of Counseling Psychology, 57*(3), 345–360.[*1 #]

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction with Life Scale Adapted for Children: A focus on cognitive processes. *Social Indicators Research, 100*, 37–60.[*1 *2]

Guirguis, L. M., & Chewning, B. A. (2008). Development of a measure to assess pharmacy students' beliefs about monitoring chronic diseases. *Research in Social and Administrative Pharmacy, 4*, 402–416.[*2 o]

Hansen, C. J., & Zytowski, D. G. (1979). The Kuder Occupational Interest inventory as a moderator of its predictive validity. *Educational and Psychological Measurement, 39*(1), 107–118.[*3 □]

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219–230.[*1 *2 *3]

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, Volume 1* (pp. 3–19). Washington, DC: American Psychological Association Press.[*1 *2]

James, W. (1890). *Principles of psychology*. New York: Holt.[*2]

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education/Praeger.[*1 *2 *3]

Kuder, F., Diamond, E. E., & Zytowski, D. G. (1998). Differentiation as fundamental validity for criterion-group scaled interest inventories. *Educational and Psychological Measurement, 58* (1), 38–41.[*3 □]

Lee, D. G., & Park, H. J. (2011). Cross-cultural validity of the frost multidimensional perfectionism scale in Korea. *The Counseling Psychologist, 39*(2), 320–345.[*1 #]

Locke, B. D., Buzolitz, J. S., Lei, P. W., Boswell, J. F., McAleavey, A. A., Sevig, T. D., Dowis, J. D., & Hayes, J. A. (2011). Development of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62). *Journal of Counseling Psychology, 58*(1), 97–109.[*1 #]

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.[*2]

Mak, L., & Marshall, S. K. (2004). Perceived mattering in young adults' romantic relationships. *Journal of Social and Personal Relationships, 21*, 469–486.[*2 o]

Marcus, F. M. (1991). *Mattering: Its measurement and theoretical significance*. Unpublished manuscript.[*2]

Marshall, S. K. (2001). Do I matter? Construct validation of adolescents' perceived mattering to parents and friends. *Journal of Adolescence, 24*, 473–490.[*2 o]

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37*, 113–115.[*1]

Mercer, S. H., Zeigler-Hill, V., Wallace, M., & Hayes, D. M. (2011). Development and initial validation of the Inventory of Microaggressions Against Black Individuals. *Journal of Counseling Psychology, 58*(4), 457–469.[*1 #]

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.[*1 *2 *3]

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.[*1 *2 *3]

Miller, M. J. (2010). Testing a bilinear domain-specific model of acculturation and enculturation across generational status. *Journal of Counseling Psychology, 57*(2), 179–186.[*1 #]

Miller, M. J., & Sendrowitz, K. (2011). Counseling psychology trainees' social justice interest and commitment. *Journal of Counseling Psychology, 58*(2), 159–169.[*1 #]

Mohr, J. J., & Kendra, M. S. (2011). Revision and extension of a multidimensional measure of sexual minority identity: The Lesbian, Gay, and Bisexual Identity Scale. *Journal of Counseling Psychology, 58*(2), 234–245.[*1 #]

Myers, J. E., & Bechtel, A. (2004). Stress, wellness, and mattering among cadets at West Point: Factors affecting a fit and healthy force. *Military Medicine, 169*, 475–482.[*2]

Nadal, K. L. (2011). The Racial and Ethnic Microaggressions Scale (REMS): Construction, reliability, and validity. *Journal of Counseling Psychology, 58*(4), 470–480.[*1 #]

Nafziger, D. H., & Helms, S. T. (1974). Cluster analyses of interest inventory scales as tests of Holland's occupational classification. *Journal of Applied Psychology, 59*(3), 344–353.[*3 □]

Nugent, F. A., & Jones, K. D. (2005). *Introduction to the profession of counseling* (4th ed.). Hoboken: Wiley.*1

Pinterits, E. J., Poteat, V. P., & Spanierman, L. B. (2009). The White Privilege Attitudes Scale: Development and initial validation. *Journal of Counseling Psychology, 56*(3), 417–429.*1 #

Powers, A., Myers, J., Tingle, L., & Powers, J. (2004). Wellness, perceived stress, mattering, and marital satisfaction among first year medical residents and their spouses: Implications for education and counseling. *The Family Journal, 12*, 26–36.*2

Reeve, B. B., Willis, G., Shariff-Marco, S. N., Breen, N., Williams, D. R., & Gee, G. C., et al. (2011). Comparing cognitive interviewing and psychometric methods to evaluate a racial/ethnic discrimination scale. *Field Methods, 23*, 397–419.*2

Rosenberg, M. (1989). *Society and the adolescent self-image* (rev. ed.). Middletown: Wesleyan University Press.*2

Rosenberg, M., & McCullough, B. C. (1981). Mattering: Inferred significance and mental health among adolescents. *Research in Community Mental Health, 2*, 163–182.*2

Rottinghaus, P. J., Coon, K. L., Gaffey, A. R., & Zytowski, D. G. (2007). Thirty-year stability and predictive validity of vocational interests. *Journal of Career Assessment, 15*(1), 5–22.*3 □

Salahuddin, N. M., & O'Brien, K. M. (2011). Challenges and resilience in the lives of urban, multiracial adults: An instrument development study. *Journal of Counseling Psychology, 58* (4), 494–507.*1 #

Savickas, M. L., Taber, B. J., & Spokane, A. R. (2002). Convergent and discriminant validity of five interest inventories. *Journal of Vocational Behavior, 61*(1), 139–184.*3 □

Schlossberg, N. K. (1990). *The mattering scales for adult students in postsecondary education*. Center for Adult Learning and Educational Credentials. Washington, DC: American Council on Education.

Schwartz, S. J., & Finley, G. E. (2010). Troubled ruminations about parents: Conceptualization and validation with emerging adults. *Journal of Counseling & Development, 88*(1), 80–91.*1 #

Sifford, A., Ng, K. M., & Wang, C. (2009). Further validation of the Psychosocial Costs of Racism to Whites Scale on a sample of university students in the southeastern United States. *Journal of Counseling Psychology, 56*(4), 585–589.*1 #

Strong, E. K. (1943). *Vocational interests of men and women*. Stanford: Stanford University Press.*3

Tovar, E., Simon, M. A., & Lee, H. B. (2009). Development and validation of the college mattering inventory with diverse urban college students. *Measurement and Evaluation in Counseling and Development, 42*, 154–178.*2 o

Vogel, D. L., Wade, N. G., & Ascheman, P. L. (2009). Measuring perceptions of stigmatization by others for seeking psychological help: Reliability and validity of a new stigma scale with college students. *Journal of Counseling Psychology, 56*(2), 301–308.*1 #

Wang, K. T., Yuen, M., & Slaney, R. B. (2009). Perfectionism, depression, loneliness, and life satisfaction a study of high school Students in Hong Kong. *The Counseling Psychologist, 37* (2), 249–274.*1 #

Wei, M., Alvarez, A. N., Ku, T. Y., Russell, D. W., & Bonett, D. G. (2010). Development and validation of a Coping with Discrimination Scale: Factor structure, reliability, and validity. *Journal of Counseling Psychology, 57*(3), 328–344.*1 #

Yoo, H. C., Burrola, K. S., & Steger, M. F. (2010). A preliminary report on a new measure: Internalization of the Model Minority Myth Measure (IM-4) and its psychological correlates among Asian American college students. *Journal of Counseling Psychology, 57*(1), 114–127.*1 #

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.*1 *2 *3

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: Information Age Publishing.*1 *2 *3

Zytowski, D. G. (1972a). A concurrent test of accuracy-of-classification for the Strong Vocational Interest and Kuder Occupational Interest Survey. *Journal of Vocational Behavior, 2*(3), 245–250.[*3] □

Zytowski, D. G. (1972b). Equivalence of the Kuder Occupational Interest Survey and the Strong Vocational Interest Blank revisited. *Journal of Applied Psychology, 56*(2), 184–185.[*3] □

Zytowski, D. G. (1976). Predictive validity of the Kuder Occupational Interest Survey: A 12-to 19-year follow-up. *Journal of Counseling Psychology, 23*(3), 221–233.[*3] □

Zytowski, D. G. (1992). Three generations: The continuing evolution of Frederic Kuder's interest inventories. *Journal of Counselling Development, 71*(2), 245–248.[*3] □

Zytowski, D. G., & Laing, J. (1978). Validity of other-gender-normed scales on the Kuder Occupational Interest Survey. *Journal of Counseling Psychology, 25*(3), 205–209.[*3] □

Part III
Psychology and Education

Chapter 6
What Counts as Evidence: A Review of Validity Studies in *Educational and Psychological Measurement*

Benjamin R. Shear and Bruno D. Zumbo

Introduction

When using a test or any other measure, the first question we must always ask ourselves is: what does the test score mean? We often glide over this question in our daily lives. When we hear about the latest results from the Program for International Student Assessment (OECD 2013), we make inferences about relative international student achievement. When a patient obtains a high score on a depression inventory, we make a judgment about possible depression (e.g., Low and Hubley 2006). These are often reasonable inferences and assumptions. Somewhere along the line, however, we hope that well-qualified and well-intentioned people have investigated these interpretations to provide evidence that, in fact, these are justifiable inferences.

In the technical measurement and testing field, this process is known as validation. In general, when a careful investigation provides evidence to support interpretations and decisions made from test[1] scores, we say those interpretations and decisions are valid (American Educational Research Association et al. 1999). Not surprisingly for such a complex and often high-stakes process, however, there is disagreement among experts as to what counts as evidence, how that evidence

[1] We will use the terms "test," "measure," "scale," and "assessment" interchangeably in this paper, although each may be used to refer to somewhat different procedures in other contexts.

B.R. Shear
Graduate School of Education, Stanford University, 485 Lasuen Mall,
Stanford, CA 94305, USA
e-mail: bshear@stanford.edu

B.D. Zumbo, Ph.D. (✉)
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_6,

ought to be gathered and evaluated, and even what it means to say something is "valid" (e.g., Lissitz 2009). In light of emerging and sometimes conflicting theories of validity and validation, one may wonder how the practice of test validation reflects these theoretical discussions, and how it may have changed over time. In today's scholarly literature, what is taken as sufficient evidence to support the inferences made from test scores? What counts as evidence of validity in practice? This chapter empirically surveys validation practices in education and psychology, as documented in the "Validity Studies" section of the journal *Educational and Psychological Measurement*, across two decades and viewed from the perspective of modern validity theory.

The section "Validity and validation" briefly traces the history of validity theory by exploring nine different concepts of validity. The section "Prior empirical studies" summarizes previous research that has empirically surveyed validation practices. Section "Methods" presents the methods used to conduct a small-scale systematic review of current and past validation practices. Sections "Results" and "Discussion" summarize and discuss the results.

Validity and Validation

Validity theorists have highlighted the important distinction between validity and validation (Borsboom et al. 2004; Zumbo 2007, 2009). While validity is the property or relationship we are trying to judge, validation is an activity geared towards understanding and making that judgment. Zumbo (2009) reminds us of the importance that a guiding rationale (i.e., validity) must play in selecting and applying appropriate analyses (i.e., validation), while Borsboom et al. note that failing to distinguish between validity and validation can lead to conceptual and methodological confusion. These authors are highlighting the importance of having a clear concept of validity, which can then be used to guide the use of validation methods. With this point in mind, we briefly review some prominent conceptions of validity from the past century, as outlined by Zumbo (2010). To better understand the interplay between validity and validation, we explore the validation methods implied by each definition. Figure 6.1 outlines the time periods and nine conceptualizations of validity we will consider. In so doing we do not wish to suggest that these are distinct periods with a linear progression and evolution with an end point, but rather overlapping time periods with each of the conceptualizations continuing, to some extent, today. In the empirical examination of validity evidence to follow, we will search for these different definitions of validity.

Perhaps the first institutionalized or formally stated definition of validity dates back to the 1920s (Michell 2009). Summarizing the work of a committee on psychological testing, Courtis (1921) presents the following definitions for validity and reliability: "Two of the most important types of problems in measurement are those connected with the determination of what a test measures, and of how consistently it measures. The first should be called the problem of validity, the

Period Introduced and Developed	Concept of validity	Implied validation methods
Early 1900s	A test is valid if it measures what it is supposed to.[1]	No single implied method.
1920s-1930s	Validity is about establishing whether a test is a good predictive device or short-hand (criterion validity).	Correlation with a criterion.
1930s-late 1960s	There are multiple "types" of validity.	Depends upon the type.
1950s-1960s	Validity is about evaluating the logical empiricist influenced "nomological network" in "construct validity."[2]	Empirically establishing the nomological network.
1970s-late 1990s	Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.[3]	Multiple sources of evidence used to provide a sound scientific basis for score interpretation.
1980s-present	Construct validity is a universal and interactive system of evidence; emphasizing construct representation and nomothetic span.[4]	Formal cognitive modeling and correlational techniques, among others.
2000-present	A test is valid for measuring an attribute if and only if the attribute exists and variations in the attribute causally produce variations in the outcomes of the measurement procedure.[5]	Formal cognitive modeling, among others.
	Validity is focused on test development and internal characteristics, such as content representation.[6]	Content validation and reliability analysis methods.
	Validity is having a contextualized and pragmatic explanation for variation in test scores.[7]	Developing and testing the explanation; multi-level cognitive and statistical modeling, among others.

Fig. 6.1 Concepts of validity and implied methods of validation reviewed (*Notes*: Relevant citations for each definition are indicated by superscripts and described in the main text. 1: Courtis (1921); Buckingham (1921). 2: Cronbach and Meehl (1955). 3: Messick (1989). 4: Embretson (1983, 2007). 5: Borsboom et al. (2004, 2009). 6: Lissitz and Samuelsen (2007). 7: Zumbo (2007, 2009))

second, the problem of reliability" (Courtis 1921, p. 80). Writing about intelligence testing, Buckingham (1921) presented provided the following concept of validity:

> There is at least as great a need for determining the validity of intelligence tests. By validity I mean the extent to which they measure what they purport to measure. If for educational purposes we define intelligence as the ability to learn, the validity of an intelligence test is the extent to which it measures ability to learn. In a very real sense, validity is more important than reliability. No one, for instance, is interested in the consistency of the results of a test which fails to measure the thing it is designed to measure. Such a test would merely be consistently valueless. (p. 274)

First, note that both definitions clearly view validity as a property of a particular test, rather than a test score or inference. Second, no particular process of validation necessarily follows from these definitions. Both authors, however, go on to recommend judging validity at least in part by considering the test scores' associations with other variables. These recommendations foreshadow what was to become another widely accepted definition of validity.

As researchers wrestled with ways to determine "if a test measures what it is supposed to," test scores also came to be seen in an increasingly behavioral light. Angoff (1988) describes validity and validation in the first half of the twentieth century as primarily empirical, and possibly even "atheoretic" (pp. 19–20). We will not debate the question of whether any judgment or procedure can be completely "atheoretic," but Angoff's point echoes the distinction between validity and validation raised at the beginning of this section (Borsboom et al. 2004; Zumbo 2009).

It appears, therefore, that in the first half of the twentieth century, test scores were taken to be signs or predictive devices for some future or alternative behavior. Describing the concept of validity that arose from these circumstances, Angoff (1988) writes:

> Consistent with other writers at that time, Bingham defined validity in purely operational terms, as simply the correlation of scores on a test with "some other objective measure of that which the test is used to measure" (Bingham 1937, p. 214). Guilford defined validity similarly: "In a very general sense, a test is valid for anything with which it correlates" (Guilford 1946, p. 429). (p. 20)

This definition of validity implies a particular approach to validation, albeit a relatively narrow one: correlating test scores with a criterion. These criterion measures often tended to be predictions of future behaviors or outcomes, such as performance on the job or in college (Angoff 1988). This gave rise to the notion of "criterion validity." As limitations of this narrow approach became clear, an increasing array of "validities" emerged throughout the 1940s and 1950s, including a distinction between predictive and concurrent criterion validity as well as content validity (Angoff 1988; Hubley and Zumbo 1996). These were often considered different "types" of validity, and were driven primarily by the validation methods used rather than by a theoretical framework of validity. The articulation of a new definition of validity, referred to as "construct validity" (Cronbach and Meehl 1955), was a critical point in the history of validity theory.

Although initially introduced as a fourth "type" of validity, construct validity brought with it a shift in perspective as well. Construct validity was initially intended to provide guidance for evaluating test score interpretations when no adequate criterion or content definition was available. Using the philosophical and scientific principles of logical empiricism (Zumbo 2010), Cronbach and Meehl (1955) outlined an approach to articulating and testing a proposed nomological network, of which test scores were one observable result.

Cronbach and Meehl's (1955) description of construct validity is not easily distinguished as either a definition of validity or a process of validation.[2] Cronbach and Meehl clearly articulated, for example, how one might go about gathering evidence during the process of validation. But they also emphasized that, "Construct validity is not to be identified solely by particular investigative procedures, but by the orientation of the investigator" (Cronbach and Meehl 1955, p. 282). Along with this came an emphasis that validity and validation were about evaluating proposed interpretations of test scores, rather than a test itself. This remains a fundamental tenet of modern validity theory (Sireci 2009). Despite this call for a holistic framework of scientific inquiry, validity remained a fragmented concept, and the type of validity one demonstrated was most often a product of the method used to document validity (Hubley and Zumbo 1996).

[2] We can note, for example, that they variously refer to both "construct validity" (p. 281) and "construct validation" (p. 299).

In an attempt to bring together these various strands of validity and validation, Messick (1989) provided the following definition of validity: "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 or other modes of assessment" (p. 13, emphasis original). This is echoed in the definition of validity presented by the AERA, APA, and NCME (1999) *Standards for Educational and Psychological Testing* (henceforth, the *Standards*), which Messick helped to draft: "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests...The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (AERA et al. 1999, p. 9). While this definition of validity does not entail a single approach to validation, three widely accepted guiding tenets are that, (a) numerous sources of evidence can contribute to a judgment of validity, (b) validity is a matter of degree rather than all or none and, (c) one validates particular uses and interpretations of test scores, rather than a test itself.

In addition to this institutionalized definition of validity, Zumbo (2010) highlights four additional concepts of validity that have been proposed recently as debate surrounding the meaning of validity and validation continues. Embretson (1983, 2007) presents a view of validity as a "universal and interactive system" (2007, p. 452). This conception of validity draws heavily on Embretson's (1983) own notion of construct representation versus nomothetic span; the former dealing largely with cognitive processes and modeling, and the latter with observed relationships between test scores and other variables. This concept of validity places substantial emphasis on modeling cognitive processes and internal test characteristics, while also providing a framework for integrating multiple forms of evidence.

Lissitz and Samuelsen (2007), in an explicit departure from modern, unified approaches to validity, define validity as related solely to internal test characteristics. They write: "Together, we suggest that these essentially internal characteristics (reliability and content validity) be called the *internal validity* of the test, and all other characteristics be considered essentially external matters" (p. 446). Part of their aim is to outline a concept of validity with more clearly developed and practical methods of validation. It is clear that their conception is well-suited to modern methods of content validation, cognitive modeling, and reliability analysis (p. 445). While they recognize the importance of additional sources of evidence, they seem to consider these as distinct from a determination of validity.

Borsboom et al. (2004, 2009) have also advocated a radically different definition of validity. This definition states that: "a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (p. 1061). This is in stark contrast to recent descriptions of validity as a property of test scores or inferences, rather than of tests themselves. Borsboom et al. suggest methods of validation that include a strong emphasis on stating formal cognitive theories, developing tests from these theories, and empirically investigating "response behavior" (p. 1068).

Building upon the construct validity framework (e.g., Cronbach and Meehl 1955; Cronbach 1971; Messick 1989), Zumbo (2007, 2009, 2010) has described validity as a contextualized and pragmatic form of explanation. In this framework, validity is an emergent property, one that arises when an inference to the best explanation for observed test score variation supports proposed inferences and interpretations. Such a property depends upon the context of measurement as well as the context of interpretation and explanation. This definition provides a unifying framework that can integrate numerous modern psychometric and statistical methods, although it does not give precedence to any particular method. Zumbo (2007, 2009) has illustrated, for example, that this view provides a theoretical way to incorporate the context of measurement (such as social psychological aspects) and increasingly complex data structures (such as measurement heterogeneity and multi-level structures), as well as the cognitive modeling associated with definitions described above.

Another influential development in validity theory is the articulation of an argument-based approach to validation (Cronbach 1988; Kane 1992, 2006, 2013; Shepard 1993). However, we do not include this in the overview of concepts of validity because it does not derive from or require a particular definition of validity, and can instead be used as a tool to support validation efforts guided by different definitions of validity. As Kane notes, the argument-based approach provides a "methodology or technology for validation" (Kane 2004, p. 136) rather than a definition of validity. We should note that Kane developed this method initially to support the investigation of construct validity as it is described by Messick (1989) and the *Standards*, and it is consistent with those views of validity.

Over time the concept of validity has evolved, as have the methods of validation appropriate for those conceptions. Importantly, we do not believe that definitions of validity have evolved to a higher or better state; instead they have changed as our social constructions and worldviews have changed, which may be better described as evolution through "natural drift" (Varela et al. 1991). As conceptions of validity grow more expansive, so too do the methods of validation that are entailed. Recently, some theories have sought to re-orient our primary focus (Embretson 2007; Zumbo 2009), while others have sought to re-orient and narrow our focus (e.g., Borsboom et al. 2004; Lissitz and Samuelsen 2007). This theoretical history is one way to explore the concept of validity. Another approach is to ask the empirical question, how have test validation practices changed over time? Such a question could study both conceptions of validity and methods of validation. Below we briefly review previous empirical studies designed to address these questions.

Prior Empirical Studies

Hogan and Agnello (2004) cite Ward et al. (1975) as one of the earliest studies to document and evaluate reliability and validity evidence as presented in published measurement research. Although the judges in these early studies considered the

reliability and validity evidence to be an important part of their evaluations, there is no specific data about how these qualities impacted their decisions. Since the early 1990s, a number of studies have investigated the reliability and validity evidence provided in support of both published and unpublished educational and psychological measures. Some of these studies looked only at reliability evidence (Hogan et al. 2000; Willson 1980), while others include evaluations of validity evidence (Cizek et al. 2008, 2010; Hogan and Agnello 2004; Jonson and Plake 1998; Meier and Davis 1990; Qualls and Moss 1996; Whittington 1998).[3]

Three studies addressed the question of whether validity evidence was reported, without providing detailed information about the nature of such evidence. Meier and Davis (1990) looked at 3 years' worth of measures used in articles from the *Journal of Counseling Psychology*: 175 articles across 1967, 1977, and 1987. They found very low rates of reporting validity evidence: evidence based on a secondary source ranged from 4 % in 1977 to 5 % in 1987; evidence based on the primary sample ranged from 2 % in 1967 to just 1 % in 1987. They found slightly higher rates of reporting reliability evidence. Qualls and Moss (1996) compared research in the 1992 volumes of APA journals with published professional standards, including the 1985 edition of the *Standards*. Their sample consisted of 622 studies using 2,167 measures. Focusing on paper and pencil measures, they found that 47.5 % of the instruments used did not include either reliability or validity evidence. Approximately 31.7 % of the articles provided validity evidence, with 74.9 % of these articles providing construct validity evidence (as opposed to either criterion or content-related evidence). Reliability evidence was reported for 41 % of the measures, and the majority (90.4 %) were estimates of internal consistency.

A subsequent study of 220 articles from journals listed in Cabell's found that 94 % of articles included validity evidence to support the measures used (Whittington 1998). The comparability of this result is unclear, however, because the coding of validity evidence was much more lenient than in other studies: "For example, if the author(s) of the article cited the source for a published measure, or listed the traits measured by each scale in a personality measure, this was treated as an attempt at reporting validity" (Whittington 1998, p. 28).

These studies suggest that rates of reporting reliability and validity information have lagged behind expectations, although the situation may be worse for validity than for reliability. The authors of all three studies emphasize the lack of evidence reported more than the quality of the evidence. Since 1998, at least five studies have extended these reviews by studying both the rate and content of the validity evidence reported.

Jonson and Plake (1998) provide the most detailed account of the specific validity evidence presented, although their analysis is limited to multiple reviews

[3] More recently Wolming and Wikström (2010) and Schafer et al. (2009) have also studied validation practices. These studies are somewhat different in their orientation and fall beyond the scope of the present study. The conclusions of Wolming and Wikstrom are considered in the Discussion section.

of a single measure. The authors were interested in whether professional and theoretical conceptions of validity had shaped the practice of validation, or vice versa. To answer this question they documented the validity evidence provided for multiple editions of a single mathematics test published during the period 1937–1995. They compared this evidence to the conceptions of validity presented in the different editions of the *Standards* published during the same period. They describe their process in four steps:

> The first step was to operationalize validity requirements from five periods of validity history. The second step was to choose an achievement test with editions that spanned that history of validity. The third step was to obtain reviews of the selected test from the *Mental Measurement Yearbooks*. . .and then document the type of validity evidence that was identified in each. The final step was to compare within the respective time periods the validity evidence presented in the test reviews with the then extant standards of validity theory. (Jonson and Plake 1998, p. 739)

The authors did not find a clear relationship between the standards and observed practice; sometimes practice followed previously published standards, and other times the descriptions in the standards appeared to follow what had already been implemented in practice. They found a greater emphasis on content validity and construct validity evidence than on criterion-related validity evidence. The discussion of construct validity increased during the post-1985 *Standards* time period, suggesting that validity practice was becoming more theoretical, as Angoff (1988) claims. They did not find a clear adoption of the increasingly theoretical frameworks put forth in the concurrent versions of the *Standards*, however. They concluded that, "It was found that the test standards do seem to be influential in forming measurement professionals' overall concept of validity but are not as influential in determining the actual validity requirements that should be applied" (p. 751).

Three more recent studies also document the nature of validity evidence reported for tests (Cizek et al. 2008, 2010; Hogan and Agnello 2004). Hogan and Agnello searched the *Directory*, Volume 7, which reports on 2,078 measures presented in 36 journals from 1991 to 1995, while Cizek et al. used the 16th *Mental Measurements Yearbook* (MMY), containing information about tests published or revised from 2003 to 2005. Hogan and Agnello reviewed 696 tests in the *Directory*, and classified the types of validity evidence into eight categories. Approximately 55 % of studies reported some form of validity evidence, with 52.3 % reporting one source of validity evidence and 2.3 % reporting two sources (none reported more than two). Of the evidence reported, 92 % comprised correlations with other variables, while the remaining 8 % was a mix of group contrasts, factor analyses, and unidentified sources. As in earlier studies, they found higher rates of reporting reliability evidence, with 94 % of studies reporting an estimate of reliability. They found little content-related evidence, however, which is at odds with some previous (e.g. Jonson and Plake 1998) and subsequent (e.g. Cizek et al. 2008) findings.

Cizek et al. (2008) conducted a similar search, although they standardized their coding scheme by basing it on the sources of evidence listed in the 1999 *Standards* and noting whether validity was presented within a modern, unified framework or

not. The most commonly reported sources of evidence included construct validity (58 % of measures), concurrent criterion (50 %), and content (48 %). Cizek et al. did not elaborate on what precisely constitutes "construct validity" evidence in their study, which is not one of the sources described in the most recent *Standards*. Evidence based upon response processes (1.8 %) and consequences (2.5 %) were presented least frequently. The infrequent reference to modern validity theory and lack of evidence based on consequences led Cizek et al. (2008) to advocate major revisions to modern validity theory. In a follow-up study to document the presence of validity evidence based on test consequences, the authors surveyed a wider range of sources but found no such examples (Cizek et al. 2010).

In summary, the earliest reviews examined reporting practices at a broad level, noting whether any validity evidence was presented or not. Subsequent studies have broadened the scope of their samples, while further systematizing their coding criteria to record the nature of the evidence reported. Specific findings and proportions vary, but there is evidence to support the assertion of a disconnect between validity theory and the practice of validation, embodied in the frequency and nature of validity evidence reported, as well as the framing of that evidence.

Missing from this literature, however, is a review of research practices that are explicitly positioned as applying and presenting accepted validation methods. In other words, although it is reasonable to compare the accumulated and reported evidence described above with concurrent theory, the sources reviewed may not have been primarily focused on validation. In some cases authors reviewed substantive research articles that used quantitative measures, while in other cases authors reviewed yearbooks that synthesized a variety of sources. The present study documents the nature of validity evidence presented in primary research studies presented as using current, accepted validation practices in the field of educational and psychological measurement. Following Cizek et al. (2008), we used the 1999 *Standards* to operationalize current validity theory. We used the "Validity Studies" section published in the journal *Educational and Psychological Measurement* and dating back to 1953 (SAGE 1953) to operationalize validation practices.

Methods

Data Sources and Limitations

To answer the question, "what counts as validity evidence in practice?" we required a way to operationalize validation practices, as well as a framework for describing the evidence encountered. We used the "Validity Studies" (VS) section of the journal *Educational and Psychological Measurement* to operationalize validation practices for two reasons. First, studies published in the VS section are peer-reviewed and focus primarily on technical measurement issues. Hence these studies

represent explicitly accepted validation practices at the time of their publication, and provide an additional perspective on validation practices to those described above. Second, because the VS section was published continuously from 1953 through 2009, it provides a longitudinal data source that can be used to track changes in validation practices over a substantial period of time.

Despite these clear strengths to using the VS section, there are also some limitations that should be noted. Research published in a single journal may not generalize to all educational and psychological research, or to other fields of research. For example, it may be that the nature of the request for studies influences the type of research published in this section, so that the studies are not necessarily representative of all aspects of current validation practice. Second, there may be a publication bias, whereby the studies that are accepted for publication do not necessarily represent all validation work being conducted, either because some work is not accepted for publication or because some work (e.g., by private companies or government organizations) cannot be published. We assume that this bias would, if anything, lead published validation studies to overstate the quality of validation practices. Despite these limitations, we felt this source provided a unique and worthwhile dataset to investigate.

Systematic Review Methodology

Sample

We used the following systematic procedure to select studies for this review. First, all articles published in the VS section (and available on-line) from 1960 to 1969 (Volumes 20–29; n = 265 studies) and from 2000 to 2009 (Volumes 60–69; n = 293 studies) were tabulated in chronological order. The decade 1960–1969 was selected to represent validation practices after the seminal work on construct validity (Cronbach and Meehl 1955) had been introduced, but when discussions of different "types" of validity was still common in the theoretical literature and was reflected in the APA *Standards* (Sireci 2009). The decade 2000–2009 was selected to represent current validation practices. We used a pseudo-random sampling function to select 20 studies from each population. Studies were screened to ensure that they included primary validity evidence before being included in the final sample. According to this criteria, one study was eliminated from the 2000 to 2009 sample, as it provided only meta-analytic reliability data; a 21st random index was used to select a replacement for this study.

Coding

We coded the validity evidence presented in each study using definitions presented in the 1999 *Standards* (AERA et al. 1999), following a similar procedure used in

earlier studies (e.g., Cizek et al. 2008). The *Standards* describe five sources of validity evidence: (a) evidence based on test content, (b) evidence based on response processes, (c) evidence based on internal structure, (d) evidence based on relations to other variables, and (e) evidence based on the consequences of testing. Evidence based on test content may include "logical or empirical analyses" (AERA et al. 1999) of the adequacy and appropriateness of a test's content, often conducted by subject matter experts (Sireci 1998). Evidence based on response processes includes theoretical and empirical analyses of the processes engaged in by test-takers, such as think-aloud protocols. Evidence based on a test's internal structure includes analyses to evaluate whether the structure underlying test items or components conforms to theoretical expectations, and may include factor analytic or item response theory (IRT) modeling. Evidence based on a test score's relationship with external variables, such as scores on other tests or subsequent observed performances, includes the type of predictive/concurrent criterion studies described above as well as convergent or discriminant correlational evidence. Evidence based on the consequences of testing involves using data about the consequences of test use to evaluate the proposed interpretations or decisions based on test scores. For each study, we coded which (if any) of these five sources of evidence was reported.

We also coded whether an estimate of test score reliability was presented, whether a new or previously used measure was being studied, and whether the authors cited a guiding theoretical framework of validity and validation (e.g. Messick 1989, the *Standards*, or another framework from those discussed above).

Coding was based on our classification of the evidence provided rather than the classification of the authors (if provided). For example if a study presented a factor analysis of a single, unidimensional measure as "convergent validity" evidence (considered relations to other variables in the *Standards*), it would instead be coded as evidence based on the internal structure of the test. All relevant evidence was coded, even if it was not referred to explicitly as "validity" evidence. We only coded primary evidence gathered and presented in the study, not evidence in prior studies that were cited.

The studies published from 1960 to 1969 were often short and included results from a single analysis, such as a regression model. A single reader coded these studies. The studies published from 2000 to 2009 were longer and more complex, often presenting multiple sources of validity evidence. To increase the reliability of our coding results, two independent readers coded each study from 2000 to 2009. Any disagreements were resolved by consensus to achieve the final coding for each study. Although only the broad category of evidence is presented here (e.g. relations to other variables or content-related evidence), the readers also made note of the particular evidence or analysis presented. The only consistent source of disagreement revolved around whether an external variable might qualify as a "criterion," as opposed to a "convergent" measure. Note, however, that either classification led to the same result in our coding scheme. It is also worth pointing out, as discussed above, that the greatest challenge in criterion-related validity

studies is identifying and measuring an adequate criterion. This disagreement may reflect that larger challenge. In summary, although the procedure was designed to yield systematic, reproducible results, there was the possibility for ambiguity in how to best classify the complex combination of evidence presented. The results represent the best judgment of the participating researchers, rather than an objective or absolute truth.

Results

The results are summarized according to the three primary research questions: (1) What types of validity evidence were reported? (2) Which definitions or frameworks of validity were used? And (3) How have these changed over time? To answer the first question, we present descriptive frequencies about the sources of evidence reported across the two decades. To answer the second question, we examine whether the authors provided an explicit definition of validity or cited a guiding theoretical framework such as the most recent edition of the *Standards*, Messick's (1989) overview, or another definition described above. Finally, to compare how practices may have evolved over time, we compare the findings for questions 1 and 2 across the two decades studied.

The frequency of reporting each source of evidence is presented in Fig. 6.2 and Table 6.1 for each decade. The sample of articles in the 1960s had a mean length of 6.25 pages ($SD = 3.61$) while articles in the 2000s had a mean length of 20.3 pages ($SD = 4.03$). Although not presented, the number of citations per article also increased substantially, so that at least one page of additional length is due to increased references.

In the 1960s sample, the majority (85 %) of the tests were pre-established tests and measures, rather than studies of newly developed tests. All but one study presented evidence based on a test score's relationships to other variables (ROV), and this was the most common source of evidence. The only study not presenting ROV evidence examined the effect of social desirability response sets. Although it was difficult to determine at times whether an external variable might constitute a "criterion" or only a convergent measure, more than half of the studies could be considered as presenting criterion-related validity evidence. Most often these were predictive criterion studies evaluating whether test scores could predict later outcomes such as grade point average. Two studies (10 %) reported evidence based on response processes of respondents, and one study (5 %) provided evidence based on the test's internal structure (excluding estimates of coefficient alpha, which were coded as reliability evidence). Evidence based on response processes included analysis of potential social desirability response sets and the impact of time taken on an examinee's scores. No studies systematically explored content-related evidence or evidence based on the consequences of test use. Four studies (20 %) reported an estimate of reliability.

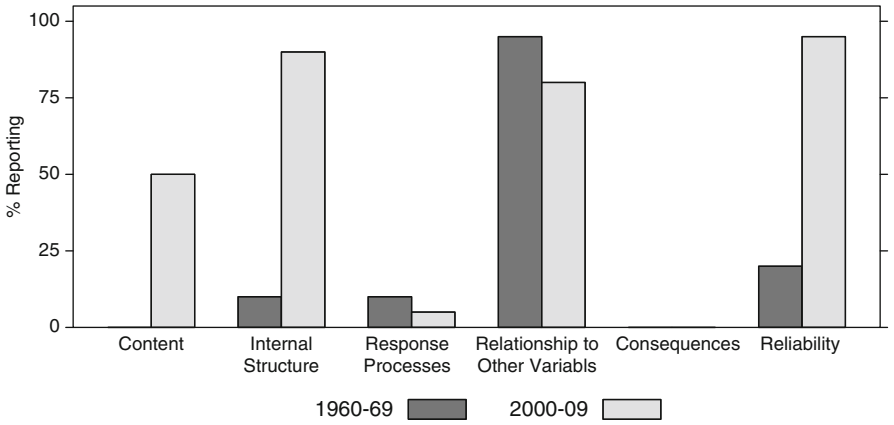

Fig. 6.2 Percent of studies reporting each source of validity evidence by decade

Table 6.1 Percent of studies reporting each source of validity evidence by decade

	1960–1969		2000–2009	
Source of evidence	%	N	%	N
Content	0 %	(0)	50 %	(10)
Response processes	10 %	(2)	5 %	(1)
Internal structure	10 %	(2)	90 %	(18)
Relationship to other variables	95 %	(19)	80 %	(16)
Consequences	0 %	(0)	0 %	(0)
Reliability	20 %	(4)	95 %	(19)
New measures	15 %	(3)	60 %	(12)
Average number of pages (SD)	6.25 (3.6)		20.3 (4.0)	

In the 2000–2009 sample, 12 (60 %) studies evaluated either new or revised versions of previously published measures. Evidence based on a test's internal structure was the most common source of evidence, used in 18 (90 %) of the studies. The two studies not citing internal structure evidence included a differential predictive criterion study of a large-scale achievement test and a study evaluating the internal consistency (reliability) and convergent/discriminant correlations of a personality inventory across two cultures. The evidence based on internal structure included factor analyses, item response theory (IRT) analyses, or some combination of the two. Factor analyses were the most common form of internal structure evidence. ROV evidence was presented in 16 (80 %) of the studies. Convergent or discriminant correlations were most frequent, with some evidence comparing different known groups, and the least emphasis on predictive criterion studies. Internal structure and ROV evidence sometimes overlapped; for example when authors conducted tests of measurement invariance, comparing the fit of confirmatory factor analysis (CFA) models across different groups.

Exactly half of the studies reported some form of content validity evidence. Importantly, even though half of the current studies included some sort of content-related

evidence, they did not always employ systematic procedures of content validation that have been advocated by methodologists, such as measures of agreement between subject matter experts (Sireci 1998). Instead, content-based evidence tended to be a description of translation procedures or a description of the process used to generate items based on the construct under study. There was little reporting of evidence based on response processes (only one study, or 5 %) and no systematic analysis of consequences. The study reporting evidence based on response processes developed a new set of tasks and regressed IRT-based difficulty estimates and response time for each task on relevant task characteristics to gain insight into how they impacted responses. All but one study (95 %) included an estimate of reliability, most often Cronbach's coefficient alpha.

The most noticeable changes across the two decades were in the proportions of studies reporting evidence based on test content, internal structure, and reliability estimates. As noted above, the breakdown of evidence used for ROV also changed, although this category remained prominent across both decades. The shift from criterion-related evidence to convergent and discriminant evidence is in keeping with theoretical descriptions of validity. The increase of systematic evidence based on test content was clear; no studies in the 1960s systematically evaluated test content, compared to approximately half in the past decade. This may be due in part to the fact that many tests studied in the 1960s were pre-existing, and content analyses had already been conducted. The increase in evidence based on internal structure is noticeable. Nearly every study in the 2000s (90 %) included some form of this evidence, while only two studies in the 1960s did. In addition, the variety of evidence for any given study also increased. In the 1960s the average number of sources was 1.15 (mode = 1, median = 1) while in the 2000s the average was 2.25 (mode = 2, median = 2). Finally, it appears that reporting an estimate of reliability has become an accepted norm given the increase in reporting these estimates (from 20 to 95 %).

No clear theoretical frameworks of validity were provided in the 1960s and few explicit definitions were provided. The language seemed to imply that validity was a property of the tests being validated, and there was no reference to theoretical articles of the period (e.g., earlier editions of the *Standards* or Cronbach and Meehl 1955). The findings were similar for the most recent decade. Rarely was an explicit definition of validity given, although it appeared that the language was tending towards discussing the validity of scores and inferences.

Two articles (10 %) from the 2000s cited a theoretical or guiding framework of validity: one study referenced the 1999 *Standards*, and another applied Messick's (1995) framework to organize their validity evidence. The study citing the *Standards* noted that new validity evidence is required for novel uses or revised versions of a measure, and investigated a revised version of a measure. The study using Messick's framework explicitly gathered and organized evidence based on four of the six aspects of construct validity described by Messick. These included evidence based on the substantive, content, generalizability, and external aspects of validity, but not based on the structural or consequential aspects. Note that these categories do not correspond directly to the categories described in the *Standards*. The

evidence in this study was coded as including evidence based on test content, internal structure, and ROV.

Discussion

This chapter reviewed multiple concepts of validity from the theoretical literature and then reviewed current (2000–2009) and past (1960–1969) validation practices in educational and psychological measurement research. While prior reviews have documented validation practices, none have focused on studies that were presented explicitly as validation studies in the technical measurement literature. We found that the statistical and psychometric complexity of validation practices have increased over time, but key sources of validity evidence remain under-represented and theoretical concepts of validity such as those reviewed above are rarely used to explicitly guide the validation process.

Comparison to Prior Studies

The study by Cizek et al. (2008) was most comparable methodologically to our review of validity studies in the 2000–2009 decade. Although Cizek et al. coded internal structure evidence differently, we found similar patterns in the rate of reporting across other sources of evidence. This included higher rates of reporting evidence based on ROV and test content, and lower rates of reporting evidence based on response processes or the consequences of test use.

Results from the Hogan and Agnello (2004) sample from the 1990s appeared more similar to our sample from the 1960s. This included evidence based primarily on correlations with other variables, and little evidence based on test content or internal structure analysis. It is difficult to know what the source of these differences is, because the time period, sample, and coding scheme differ from those used here. For example, of the subsample of studies published in *EPM* that Hogan and Agnello reviewed, only about 60 % were listed as including any validity evidence, suggesting that many of the studies may not have been from the Validity Studies section. Differences could also be due to taking each test as the unit of analysis, rather than the published study (as done here). Jonson and Plake's (1998) finding that content-related evidence was more common than evidence based on ROV was not seen here in either time period. This pattern of results may provide support for the assertion that validity evidence is selectively documented based on what is most readily available (Hubley and Zumbo 1996; Kane 2004), but also could be particular to the single test studied by Jonson and Plake. Because there were Validity Studies published in *EPM* during both of these time periods, a future study extending the review presented here to those additional time periods could provide additional insight into these trends and differences.

Our findings generally support previous assertions of a disconnect between validity theory and the practice of validation (Hogan and Agnello 2004; Hubley and Zumbo 1996; Kane 2004; Messick 1988; Shepard 1993; Wolming and Wikström 2010). This is reflected in both the imbalance of the sources of evidence presented (with little or no evidence based on response processes or consequences of test use) and in the lack of explicit reference to theoretical frameworks of validity theory, including the *Standards*. As noted above, we cannot necessarily generalize these trends and explanations to all validation practices. However, the consistency of the findings here and in prior reviews suggests that these results reflect wider trends in validation research.

Possible Explanations and Implications

Theorists have proposed different explanations for the lack of reference to validity theory in practice. Some have proposed, for example, that the current unified theory of construct validity, as described by Messick (1989) and in the *Standards*, requires unattainable or unrealistic goals (Chapelle et al. 2010; Lissitz and Samuelsen 2007; Moss 2007; Shepard 1993). As a result, "the sense that the task is insurmountable allows practitioners to think that a little bit of evidence of whatever type will suffice" (Shepard 1993, p. 429). This explanation is generally consistent with our findings. An alternative explanation is that the *Standards* and other descriptions of current validity theory simply lack practical guidance (Kane 2004, 2006). Hence there is growing support for following an argument-based approach to validation, which is intended to bridge this gap by providing a methodology for prioritizing and evaluating the validity evidence that needs to be collected (Cronbach 1988; Kane 1992, 2006; Shepard 1993; Sireci 2009). In addition, other concepts of validity intended to be more useful in guiding validation research have appeared and were described above (e.g., Lissitz and Samuelsen 2007). This explanation is less consistent with our findings because we did not see uptake of these theories, although the theories and frameworks may be too recent to show up consistently in the time period we reviewed.

The shift in types of validity evidence reported could represent either practical or theoretical changes. Practically, the increased reporting of latent variable analyses and reliability estimates may partially be due to the wider availability of computational and statistical resources. Yet the drastic increase in reporting rates also suggests a subtler theoretical shift may have played a role. In a criterion-related validity framework, for example, factor analyses do not provide relevant validity evidence. In a more construct-centered framework, however, factor analyses can inform us about the theoretical constructs we are studying. Hence the changes may indicate an implicit acceptance of more theoretical concepts of validity, even if such frameworks are not explicitly used to guide validation studies (Jonson and Plake 1998; Wolming and Wikström 2010).

These explanations do not directly account for the lack of evidence based on response processes and consequences of test use. Prior authors have suggested the lack of evidence based on consequences demonstrates that "it is not possible to include consequences as a logical part of validation" (Cizek et al. 2010, p. 739), although they do not make analogous claims about response processes. Cizek et al. argued that evidence of positive or negative consequences of test use, arising either from proper or improper uses of tests, cannot logically support or refute the accuracy of a score-based interpretation and hence should be removed as a relevant aspect of validity. This is a contested claim, and the on-going debate regarding the proper role of consequences in validity and validation research is too complex to address in the present chapter (see Cizek et al. 2010; Crocker 1997; Hubley and Zumbo 2011; Kane 2013; Smith 1998). The controversy and lack of clarity surrounding this source of evidence in the theoretical literature may be another reason it is not widely studied in practice.

Haertel (2013) proposed a number of potential barriers to the evaluation of indirect (but often intended) effects of test use that may also apply to the lack of evidence based on consequences and response processes observed here. For example, the required analyses may extend beyond the disciplinary expertise of those most often conducting validation research, or the time and cost of additional data collection required for the studies may be prohibitive. Once one has collected responses to a test, it is more straightforward to compute correlations with other variables, conduct analyses of test content, or analyze the internal structure of scores than it would be to gather additional data about response processes or follow-up information about the consequences of test use.

The results of our review suggest important implications in practice. Messick (1995), for example, warned that two primary threats to the validity of score interpretations are construct underrepresentation and construct irrelevant variance. Systematic study of the response processes used by test takers or the consequences of test interpretation and use could provide key evidence needed to identify these potential threats to validity. The lack of evidence based on response processes and consequences, indicated by results here and in prior reviews, raises concern for routine interpretations of test scores.

The absence of guiding theories of validity is more troubling than the absence of any one particular concept of validity. In the absence of a clear guiding theory of validity, it is difficult to evaluate whether a particular program of validity research has accomplished its aims. This absence complicates comparisons from findings across different validity studies because they may not be trying to accomplish the same goal. It also undermines the statement in the *Standards* that validity is "the most fundamental consideration in developing and evaluating tests" (AERA et al. 1999, p. 9) because it may not be clear what exactly a concern for validity entails. There are different concepts of validity that might be used to guide validation research, including those reviewed above. Yet there is still need for greater clarification regarding specific methods of validation that can be used to evaluate test scores relative to these concepts of validity. The argument-based approach to validation provides one framework for structuring validation research,

but still seems to require a theory of validity that can serve as a guiding aim. Further developments on both of these fronts seem more important than advocating that a particular concept of validity be adopted.

We wish to make one final note regarding our discussion of validation practices: we do not take these findings as shortcomings of the individual studies reviewed. First, it would be nearly impossible to combine all relevant sources of validity evidence into any single study. And second, it is likely that in many studies the researchers were operating within a particular theory of validity, but simply did not make this explicit. Nonetheless, the patterns indicated across these studies (and those in prior reviews) does suggest important disconnects between theory and practice in validity research that may have practical consequences.

Conclusion

Over the past 50 years (1960–2009) concepts of validity have grown increasingly expansive, and methods of validation have become increasingly complex and multi-faceted. This paper has traced both histories. Current researchers commonly include more diverse evidence to support test score interpretations, with notable increases in factor analytic and content-based evidence. Yet past and present validation research has continued to leave out validity evidence based on the response processes of examinees and the consequences of test use. In addition, although researchers seem to consider more (and more complex) sources of evidence, clear theoretical bases for such practices, such as the concepts of validity described above, were not explicitly stated. We hope that validity theorists will continue to push our conceptual understandings of validity forward, while also attending to the practical implications of validation research. Meanwhile, we also hope that practitioners will be clear in stating the guiding theory or theories of validity that motivate validation studies. Others have suggested the value in disseminating exemplary validity studies, possibly with critical commentary, as a means to progress validation research (Moss 2007; Sireci 2009). The "Validity Studies" section, with a long history of publishing scholarly work in validation research could provide one potential outlet for such publications, which could further both aims. Moving forward on these issues will require substantial efforts from both validity theorists and applied researchers conducting validation studies, efforts that we believe are essential.

References

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

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

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York: Harper.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications*. Charlotte: Information Age Publishing Inc.

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology, 12*, 271–275.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13. doi:10.1111/j.1745-3992.2009.00165.x.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397–412. doi:10.1177/0013164407310130.

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*(5), 732–743. doi:10.1177/0013164410379323.

Courtis, S. A. (1921). Report of the standardization committee. *Journal of Educational Research, 4*(1), 78–90.

Crocker, L. (1997). Editorial: The great validity debate. *Educational Measurement: Issues and Practice, 16*(2), 4–4. doi:10.1111/j.1745-3992.1997.tb00584.x.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

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.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. doi:10.1037/0033-2909.93.1.179.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher, 36*(8), 449–455. doi:10.3102/0013189X07311600.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*(4), 427–438.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspective, 11*(1–2), 1–18. doi:10.1080/15366367.2013.783752.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*(5), 802–812. doi:10.1177/0013164404264120.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*(4), 523–531. doi:10.1177/00131640021970691.

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*(3), 207–215.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219–230. doi:10.1007/s11205-011-9843-4.

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*(5), 736–753. doi:10.1177/0013164498058005002.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. doi:10.1037/0033-2909.112.3.527.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspective, 2*(3), 135–170.

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi:10.1111/jedm.12000.

Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions and applications*. Charlotte: Information Age Publishing Inc.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437–448. doi:10.3102/0013189X07311286.

Low, G. D., & Hubley, A. M. (2006). Screening for depression after cardiac events using the Beck Depression Inventory-II and the Geriatric Depression Scale. *Social Indicators Research, 82*(3), 527–543. doi:10.1007/s11205-006-9049-3.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37*(1), 113–115. doi:10.1037/0022-0167.37.1.113.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale: Lawrence Erlbaum Associates.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. doi:10.1037/0003-066X.50.9.741.

Michell, J. (2009). Invalidity in validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 111–133). Charlotte: Information Age Publishing Inc.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher, 36*(8), 470–476. doi:10.3102/0013189X07311608.

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Retrieved from http://dx.doi.org/10.1787/9789264190511-en

Qualls, A. L., & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. *Educational and Psychological Measurement, 56*(2), 209–214. doi:10.1177/0013164496056002002.

SAGE. (1953). Provision for publication of validity studies involving school grades. *Educational and Psychological Measurement, 13*(1), 150–151. doi:10.1177/001316445301300116.

Schafer, W. D., Wang, J., & Wang, V. (2009). Validity in action: State assessment validity evidence for compliance with NCLB. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 173–193). Charlotte: Information Age Publishing Inc.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*(1), 405–450. doi:10.3102/0091732X019001405.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*(1), 83–117. doi:10.1023/A:1006985528729.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte: Information Age Publishing Inc.

Smith, J. K. (1998). Editorial. *Educational Measurement: Issues and Practice, 17*(2), 4–4. doi:10.1111/j.1745-3992.1998.tb00823.x.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: The MIT Press.

Ward, A. W., Hall, B. W., & Schramm, C. F. (1975). Evaluation of published educational research: A national survey. *American Educational Research Journal, 12*(2), 109–128. doi:10.3102/00028312012002109.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58*(1), 21–37. doi:10.1177/0013164498058001003.

Willson, V. L. (1980). Research techniques in AERJ Articles: 1969 to 1978. *Educational Researcher, 9*(6), 5–10. doi:10.3102/0013189X009006005.

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice, 17*(2), 117–132. doi:10.1080/09695941003693856.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (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: Information Age Publishing Inc.

Zumbo, B. D. (2010, September). *Measurement validity and validation: A meditation on where we have come from and the state of the art today*. Presented at the International conference on outcomes measurement, US National Institutes of Health, Bethesda, MD.

Chapter 7
Validity Evidence in the *Journal of Educational Psychology*: Documenting Current Practice and a Comparison with Earlier Practice

Rebecca J. Collie and Bruno D. Zumbo

Validity Evidence in the Journal of Educational Psychology Within Two Time Periods

The current *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] 1999) state that "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). According to Goodwin and Leech (2003), this view of validity is significantly different from earlier editions of the Standards (AERA et al. 1985; APA et al. 1966, 1974; APA 1954), due in large part to the evolution (Shepard 1993) or metamorphosis (Geisinger 1992) that has taken place in relation to validity theory over the past 50 years (Jonson and Plake 1998). In spite of this significant evolution, scholars (e.g., Borsboom et al. 2004; Hubley and Zumbo 1996; Messick 1988) have raised concerns over whether validation practice presented in the literature is keeping pace with this evolution in validity theory. The aim of the current study, therefore, was to document current validation practice by examining evidence presented in articles published from 2000 through 2010 in the *Journal of Educational Psychology*. In addition, the study aimed to see whether and how validation practice has changed over the past 50 years by comparing current practice with earlier practice reported in articles published from 1950 through 1960.

R.J. Collie (✉)
School of Education, University of New South Wales, Sydney, NSW 2052, Australia
e-mail: rebecca.collie@unsw.edu.au

B.D. Zumbo, Ph.D. (✉)
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_7,

The Evolution of Validity Theory

Validity theory has changed greatly from the 1950s until the present. Up to and during the early 1950s, validity was generally considered under a criterion-based model of validity (Kane 2001; however, see Rulon 1946). This view commonly involved correlating a test with an external criterion measure; if the test correlated highly with the criterion, then it was considered valid (Goodwin and Leech 2003; Jonson and Plake 1998). During this time, the validity of the test itself was the primary concern (Goodwin and Leech 2003), and the degree to which a test measured what it was purported to measure was the key to validity (Kane 2001). In addition, "test validity was a singular concept" (Jonson and Plake 1998, p. 737).

In 1952, the *Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal* (APA Committee on Test Standards) was published. It introduced "four categories of validity: predictive validity, status validity, content validity, and congruent validity" (Sireci 1998, p. 88). By 1954, when the first version of the Standards (called the *Technical Recommendations for Psychological Tests and Diagnostic Techniques*; APA) were published, the names were changed slightly to "'types' or 'attributes' of validity" (Sireci 1998, p. 88) including predictive, concurrent (previously status validity), content, and construct validity (previously congruent validity). According to Sireci (2009), it was with the publication of the 1954 Standards that "the concept of 'construct validity' was born" (p. 24). It was, however, in a seminal paper by Cronbach and Meehl (1955) that this was further elaborated. Indeed, it was in a response to that paper that Loevinger (1957) was the first to argue that construct validity is the whole of validity.

Over the decades since the first Standards (APA 1954) were published, many further changes have taken place. These have included the forgoing of *types of validity* in favor of *types of evidence* under a unitary view of validity (Jonson and Plake 1998), as well as a change in the view of validity from a property of the test to a characteristic of the test scores or inferences (Goodwin and Leech 2003). The current state of validity theory is visible in the current Standards (AERA et al. 1999), which mention five types of validity evidence. The first is evidence based on test content, which addresses "the extent to which the content of a measure represents a specified content domain" (Goodwin and Leech 2003, p. 183). The second is evidence based on response processes, which examines the processes in which participants' engage in order to respond to test questions to understand why certain responses were given by certain groups (AERA et al. 1999) and to see if they correspond with the construct being measured (Goodwin and Leech 2003). The third is evidence based on internal structure, which examines "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al. 1999, p. 13). The fourth is evidence based on relations to other variables, which analyses the "relationship of test scores to variables external to the test" (AERA et al. 1999, p. 13) and refers to traditional criterion-related validity and traditional aspects of

construct validity such as convergent and discriminant validity (Goodwin and Leech 2003). The final type is evidence based on testing consequences, which refers to the intended and unintended consequences that impact validity through construct-irrelevant variance and construct underrepresentation (AERA et al. 1999; Hubley and Zumbo 2011).

The Current Study

Despite great changes in validity theory over the past 60 years, questions remain as to whether validation practice presented in published articles has kept pace with these changes. The current study, therefore, examined articles published recently in the Journal of Educational Psychology (JEP) in an attempt to document current practice. More specifically, validation practice presented in articles published from 2000 through 2010 was examined. The current study also investigated whether and how validation practice has changed over time. To do that, comparisons were made between past and current practice to identify the degree to which practice has (or has not) changed over time. This type of examination is important given changes to the Standards (e.g., AERA et al. 1999), as well as changes in validity theory (Sireci 2009; Zumbo 2007) over the past 50 years. To do this, a second time period was also included in analyses. Validation practice reported in articles published from 1950 through 1960 was examined in addition to the contemporary articles. This earlier period was chosen for the comparison because it was a time when long held beliefs about validity (e.g., the focus on criterion-related validity) were being actively questioned and the concept of construct validity was first proposed.

In order to conduct the current study, a framework was used that stems from Cizek et al.'s (2008) study that examined validation practice reported in evaluations of measures in the 16th Mental Measurements Yearbook (Spies and Plake 2005). Two overarching research questions guided the current study:

1. What is the nature of current validation practice presented in articles as it pertains to (a) validity characteristics; (b) different sources of validity evidence; (c) number of different sources of validity evidence; and (d) justification for and types of criterion-related predictive, criterion-related concurrent, convergent, or discriminant variables?
2. To what degree has validation practice changed from articles published in 1950–1960 to those published more recently?

Data Source

To obtain data for this study, we conducted a search of articles published in JEP through the online PsycARTICLES database. Issues published between 2000 and 2010, as well as issues published between 1950 and 1960 were included in the

search. Articles that had the term *validity* or *validation* in their abstracts were included in the initial sample. For 2000–2010, this search returned 30 articles. Eleven of these articles had very little to do with validity or validation (e.g., it was mentioned once or twice, but no evidence was provided) or were using a different meaning of the word (e.g., the validity of drawings compared to real life). This left 19 articles as data for the 2000–2010 time period. For 1950–1960, the search returned 29 articles. Again, these were examined based on the content and 13 were excluded because they were theoretical articles about validation or they had very little to do with validity or validation. In total, 35 articles were utilized as data sources. The appendix contains references to all articles examined in the study.

Methods

The articles were examined for their presentation of validity evidence using a similar method to that which Cizek et al. (2008) used in their study. This involved documenting validation practice including the sources of validity evidence provided in the articles, how validity was characterized, as well as several other analyses. However, where Cizek et al. examined reviews of educational and psychological tests in the Mental Measurements Yearbook (Spies and Plake 2005), the current study examined articles on educational and psychological measures published in JEP. Therefore, in order to best answer the research questions, certain categories of examination were excluded (e.g., test type), while additional categories were added (e.g., reference to validity experts, the justification for choice of comparison variables).

Categories of Examination

The first category examined was *validity characteristics*. Four indicators were examined for validity characteristics including whether articles presented a unified or separated view of validity, made reference to validity as a property of a test or property of the inferences of a test, made reference to any version of the Standards (AERA et al. 1985, 1999; APA et al. 1966, 1974; APA 1954), and made reference to experts and/or seminal validity papers. The second category examined was *sources of validity evidence*. Indicators examined in this category were adapted from the current Standards (AERA et al. 1999) and Cizek et al.'s (2008) study. As such, seven sources of evidence were examined: evidence based on response processes, consequence of testing, test content, internal structure, predictive relations to other variables, concurrent relations to other variables, and construct. Two of these sources were further refined as per Cizek et al.'s (2008) study. First, we examined whether internal structure was reported as evidence of reliability, validity, or as reliability evidence that informs validity. Second, we analyzed four

components of construct evidence: whether the article mentioned the term *construct validity*, undertook factor analysis (FA) or structural equation modelling (SEM) to explore constructs, mentioned convergent validity, or mentioned discriminant validity. The third category examined the *number of sources* of validity evidence reported in each article. This simply involved counting how many different sources of evidence each article accurately reported. The final category examined *comparison variables*, which refer to criterion-related predictive, criterion-related concurrent, convergent, and discriminant variables. Two indicators were examined for this category: whether justification was provided for the choice of comparison variables and the types of comparison variables used in each article.

Results

The results are organized by category of examination (e.g., validity characteristics, sources of evidence). The first results reported are for current practice (i.e., articles published in 2000–2010). Following this, comparisons across the two time periods are made (i.e., 1950s versus 2000s). It is important to note that the results are based on accurate validation practice presented in the articles. This means that if articles claimed to provide a certain source of evidence, but did not follow through accurately they were not coded in that category. Among the sample, only two articles fell into this group. Both claimed to provide criterion-related predictive validity evidence, but did not use a criterion variable and/or did not use one that was measured in the future. Thus, they were not coded as presenting criterion-related predictive evidence. In addition to those articles, there were several others that accurately reported a certain source of validity evidence despite not naming it as such. In total seven articles from 1950 to 1960 and eight from 2000 to 2010 were coded as reporting a source of evidence despite not naming it in the article. For example, Mokhtari and Reichard (2002) developed a new measure called the Metacognitive Awareness of Reading Strategies Inventory by obtaining expert opinion from researchers in the field of reading strategies on the "clarity, redundancy, and readability" of the items (p. 251). Thus, although they did not call it evidence based on test content, it was classified as such given that in the current Standards (AERA et al. 1999) evidence based on test content can come from judgments by experts in the area on the relationships between items in the test and the construct.

Documenting Current Practice

The first aim of this study was to document current practice. In order to do that, the articles published between 2000 and 2010 were examined for presentation of the various categories under examination: validity characteristics, sources of validity evidence, number of sources, and comparison variables. These are discussed in turn below.

Validity Characteristics

The first category under examination concerned the characterization of validity presented in the articles (Research Question 1a). As described above, four indicators were used to assess this category. The first indicator examined whether articles presented a unitary or separated view of validity. Analyses revealed that the majority of articles reported multiple types of validity (63 %). Furthermore, 42 % of articles published at this time mentioned types of validity that are not considered in the current or previous Standards (e.g., AERA et al. 1999). These types of validity included face, internal, external, postdictive, ecological, and factorial validity. For example, d'Ailly (2003) mentioned predictive, concurrent, construct, as well as ecological validity. Several articles (26 %) referred to construct validity only and/or other types of validity as a part of construct validity. For example, "...the present study is a construct validity investigation of the [scale] by independent researchers to explore the underlying dimensions of reading motivation as assessed by the [scale]" (Watkins and Coffey 2004, p. 111). We chose to code these views of validity as unitary given that they refer to the modern view that all evidence bears upon construct validity (Sireci 2009). However, none of the recently published articles explicitly reported a unitary view of validity.

The second indicator examined whether articles referred to validity as a property of the test or as a property of the inferences of the test. The overwhelmingly majority of articles (95 %) referred to validity as a property of a test, and in some cases the property of a model (e.g., Janosz et al. 2000). For example, "...[these findings] present compelling evidence for the validity and utility of the [Academic Entitlement] scale" (Chowning and Campbell 2009, p. 994). Only one article (3 %) referred to validity as a characteristic of the inferences. This article stated, "construct validity of the interpretation of this difference as a diffusion effect was supported by comments by both the teachers themselves and by external observers" (Craven et al. 2001, p. 643).

The final two indicators for this category concerned references made to the Standards (e.g., AERA et al. 1999) and validity experts. Although the examined articles featured validation as a main part of their content, not one article made reference to the current Standards (AERA et al. 1999) or a previous version (e.g., AERA et al. 1985). More promising, however, was that six articles (32 %) made reference to one or more experts.

Sources of Validity Evidence

For the second category, articles were examined for whether they accurately presented any of the seven different sources of validity evidence: evidence based on response processes, consequences, internal structure, content, predictive

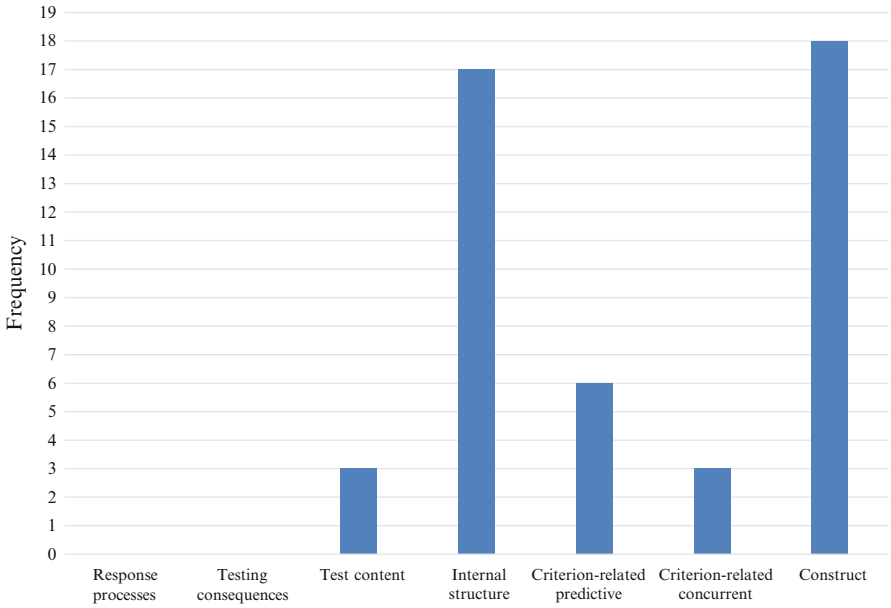

Fig. 7.1 Sources of validity evidence presented in articles published between 2000 and 2010

relations to other variables (i.e., criterion-related predictive), concurrent relations to other variables (criterion-related concurrent), and construct (Research Question 1b). The frequency with which the different sources of evidence were presented in articles published in 2000–2010 is shown in Fig. 7.1. As the figure shows, most articles reported construct evidence (95 %) and internal structure (89 %). Criterion-related predictive (32 %), criterion-related concurrent (16 %), and content evidence (16 %) were also reported in some articles. However, no articles reported evidence based on response processes or testing consequences.

Characterization of Internal Structure

In addition to coding whether articles reported evidence of internal structure, we also examined their characterization of internal structure as reliability evidence only, validity evidence only, or as reliability evidence that informs validity. Of the articles, most (89 %) reported evidence of internal structure (e.g., Cronbach's alphas). In all of these articles, internal structure was reported only as reliability. For example, "we examined the internal consistency reliability… of each of the factors constituting the given model" (Brockway et al. 2002, p. 215). In other words, none of the recent articles reported reliability as informing validity.

Components of Construct Evidence

As described above, four components of construct evidence were also documented: whether the article mentioned the term *construct validity*, undertook factor analysis (FA) or structural equation modelling (SEM) to explore constructs, provided convergent evidence, or provided discriminant evidence. All except one article reported at least one component of construct validity. The term, *construct validity*, was the most frequently reported component—84 % of articles mentioned construct validity. This was followed by reports of FA or SEM (68 %), convergent evidence (42 %) and discriminant evidence (37 %).

Number of Sources

For the third category, accurate reports of validity evidence were examined to ascertain the number of different sources of validity evidence that each article reported (Research Question 1c). Over half of the article reported one source of evidence (63 %); however, two sources (21 %, 4 articles), three sources (11 %, 2 articles), and four sources (5 %, one article) were also reported in several articles. For the article that reported four sources of evidence (i.e., Brockway et al. 2002), evidence based on test content, criterion-related predictive, criterion-related concurrent, and construct was provided.

Comparison Variables

Comparison variables were examined based on two indicators: whether articles provided justification for their choice of comparison variables and the types of comparison variables that articles used for providing evidence of validity (Research Question 1d). For the first indicator, we assessed articles on whether they provided *convincing justification*, *unconvincing justification*, or *no justification* for their choice of comparison variables. Convincing justification included a description of why the variable was chosen, whereas unconvincing arguments explained what the comparison variable was without justifying its choice or providing inaccurate reasoning for the choice of variable. Although not specifically mentioned in the current Standards (AERA et al. 1999), the provision of justification is not only good practice, but also necessary if criterion-related claims are to be upheld and clearly understood as evidence for validity.

Of the articles, 13 articles made reference to a comparison variable. Among these, eight provided no justification, two provided a convincing justification, and three provided an unconvincing justification. For one of the convincing justifications, Edwards and Schleicher (2004) explained how the variable of interest (tacit knowledge) had been correlated with the criterion variable (performance) in previous literature and what steps were needed to provide more criterion-related predictive validity evidence. For the unconvincing justifications, the articles did not

justify why the variables were chosen. For example, "Correlations between the newly developed... subscales and the published scales were examined for convergent and divergent [sic] validity" (Chowning and Campbell 2009, p. 984).

For the second indicator, we examined the types of comparison variables that articles used. Convergent (42 %), discriminant (37 %), and criterion-related predictive (32 %) variables were most frequently reported. In addition, criterion-related concurrent variables were mentioned by three articles (16 %).

Comparisons with Earlier Practice

The second aim of this study was to compare current practice with earlier practice in order to understand the degree to which validation practice has changed over time. Comparisons were made between recent articles and those published around half a century ago (from 1950 through 1960). Results are described below.

Validity Characteristics

As noted above, four indicators were used to assess validity characteristics. Table 7.1 shows the results for the four indicators by decade and in total. Before comparing the results across the two periods, it is important to present the findings from articles published in the earlier time period. As Table 7.1 shows, 69 % of articles published in 1950–1960 mentioned validity as a stand-alone concept; however, this pre-dates the contemporary unitary view of validity and refers to a singular entity—often a coefficient—that was considered proof of validity before 'types' of validity became prominent. For example, "As can be seen from a comparison of the validity coefficients for the two forms of the scale... differences in validity of the two types of response are negligible" (Neidt and Merrill 1951, p. 435). Of the other articles from that time period, 19 % mentioned multiple types of validity. These articles were published in the latter half of the 1950s and reflect the terminology change towards *types of validity* that occurred after the first Standards (APA 1954) were published. The remaining articles were unclear as to their characterization of validity. In comparing these results with those from the 2000–2010 articles, we see that the characterization of validity has changed. Contemporary articles were more likely to report 'all evidence as bearing on construct validity' (from no articles in 1950–1960 to 26 % of articles in 2000–2010). In addition, many more articles referred to multiple types of validity (from 19 % of articles in 1950–1960 to 63 % of articles in 2000–2010). However, no articles from either time period explicitly reported a unitary view of validity.

For the second indicator (i.e., validity as a property of the test versus a property of the inferences), almost all articles (94 %) published in 1950–1960 referred to validity as a property of a test. For example, "the validity of an English

Table 7.1 Characterization of validity

	1950–1960	2000–2010	Total
View of validity			
Unitary			
Explicit statement of a unitary view	0	0	0
All evidence bearing on construct validity	0	5	5
Not unitary			
Multiple types mentioned	3	12	15
Only validity mentioned	11	0	11
Unclear	2	2	4
Property of test or inferences			
Test	15	18	33
Inferences	0	1	1
Unclear	1	0	1
Reference to test standards			
Current version	0	0	0
Previous version	0	0	0
No reference made	16	19	35
Reference to experts			
Yes	2	6	8
No	14	13	27

Examination for Foreign Students was tested against the criterion of final grades. . .” (Lorge and Diamond 1954, p. 214). The only article that did not refer to validity as a property of a test was unclear in its characterization and so could not be coded. Comparing these results with current practice, we see that very little has changed. Almost all articles in both time periods refer to validity as a property of a test: 94 % of articles in 1950–1960 and 95 % of articles in 2000–2010.

The final two indicators for this category were concerned with references made to the Standards (e.g., AERA et al. 1999) and experts. Similar to the findings among recent articles, no articles published in 1950–1960 made reference to a previous version of the Standards (e.g., APA 1954). In terms of references made to experts, results suggest some change in validation practice. Only two articles (13 %) published in 1950–1960 made reference to one or more experts; however, this increased to six articles (32 %) published in 2000–2010. Experts cited included Messick (1995; cited three times), Campbell and Fiske (1959; cited twice), Cook and Campbell (1979; cited twice), Cronbach (1949; cited once), Cronbach and Meehl (1955; cited once), and Messick (1989; cited once).

Sources of Validity Evidence

The frequency with which the different sources of evidence were presented within each time period and in total is shown in Table 7.2. Among articles published in 1950–1960, there were several sources of evidence that were reported with similar

Table 7.2 Sources of validity evidence accurately reported

	1950–1960	2000–2010	Total
Response processes	0	0	0
Testing consequences	0	0	0
Test content	0	3	3
Internal structure	8	17	25
Criterion-related predictive	10	6	16
Criterion-related concurrent	7	3	10
Construct	2	18	20

Table 7.3 Characterization of internal structure

	1950–1960	2000–2010	Total
As reliability only	7	17	24
As validity only	0	0	0
As reliability and validity	1	0	1
No evidence	8	2	10

frequencies. The most frequently reported source of evidence was criterion-related predictive evidence (63 %). This was followed closely by internal structure (50 %) and criterion-related concurrent (44 %) sources of evidence. Construct evidence (13 %) was also reported, although less frequently. In comparing these results with current practice, we see that there has been an increase in reports of construct evidence (from 13 % of articles in 1950–1960 to 95 % of articles in 2000–2010) and internal structure (from 50 % of articles in 1950–1960 to 89 % of articles in 2000–2010). There was also a decrease in the two types of criterion-related evidence since the 1950s: from 63 % in 1950–1960 to 32 % in 2000–2010 for criterion-related predictive evidence and from 44 % in 1950–1960 to 16 % in 2000–2010 for criterion-related concurrent evidence.

Characterization of Internal Structure

Table 7.3 shows the characterization of internal structure across decades and in total. From 1950 to 1960, eight (50 %) of the articles provided evidence of internal structure. Of these, only one article published in 1960 reported reliability as a component of validity. In that singular case, the authors provided reliability coefficients and explained, "to further explore the question of validity of the need measures, coefficients of internal consistency of two of these measures were computed" (Uhlinger and Stephens, 1960, p. 263). All of the remaining articles that reported internal structure presented it as reliability only (i.e., not as part of validity). Comparisons across decades reveal that the vast majority of articles were and still are reporting internal structure as reliability only. However, a greater proportion of recent articles reported internal structure in any form suggesting greater uptake of this practice over time (from 50 % of articles in 1950–1960 to 89 % of articles in 2000–2010).

Table 7.4 Components of construct evidence

	1950–1960	2000–2010	Total
Construct validity mentioned	0	16	16
FA or SEM conducted	0	13	13
Convergent	2	8	10
Discriminant	0	7	7

Table 7.5 The number of different sources of validity evidence presented in articles

	1950–1960	2000–2010	Total
One source	13	12	25
Two sources	3	4	7
Three sources	0	2	2
Four sources	0	1	1

Components of Construct Evidence

Table 7.4 shows the four additional components of construct evidence that we examined and the frequency with which they were reported within the two time periods and in total. In 1950–1960, convergent evidence was the only reported component of construct evidence and it was reported in only two articles (13 %). A comparison of practice across the decades reveals that a great deal more articles reported the components of construct evidence in the recent articles and this includes reports of all four components (e.g., 84 % mentioned construct validity; see Table 7.4).

Number of Sources

Table 7.5 shows the number of sources reported by decade and in total. In 1950–1960, most articles reported only one source of evidence (81 %), with the remaining reporting two sources of evidence (19 %). Comparing these results with recent articles, we see that most articles still only reported one source of evidence. However, a greater proportion of recent articles reported multiple sources suggesting some change in practice over time (from 19 % of articles in 1950–1960 to 37 % of articles in 2000–2010).

Comparison Variables

We also assessed articles on whether they provided convincing justification, unconvincing justification, or no justification for their choice of comparison variables. All the articles published in 1950–1960 mentioned at least one comparison variable;

Table 7.6 Types of comparison variables reported

	1950–1960	2000–2010	Total
Criterion-related predictive	10	6	16
Criterion-related concurrent	7	3	10
Convergent	2	8	10
Discriminant	0	7	7

however, most provided no justification (81 %, 13 articles). Only two articles provided a convincing justification (13 %). For example,

> Despite the limitations of teachers' grades as statistical variables, we must recognize that grades are criteria in a very real sense—they are actually the principal evaluation in most school situations. It is therefore essential that tests intended as predictors be correlated with grades. (Doppelt and Wesman 1952, p. 210)

In addition, one article (6 %) provided an unconvincing argument by describing the comparison variable, but not justifying why it was chosen: "Freshman grades in college, or honor point ratio, were used as the criterion of scholastic achievement" (Holland 1959, p. 136). Comparing these results with recent articles, we see practice has only changed slightly with most articles still providing no justification (81 % of articles in 1950–1960 and 62 % of articles in 2000–2010).

Table 7.6 shows the types of comparison variables that articles reported. In 1950–1960, criterion-related predictive (63 %) and criterion-related concurrent (44 %) variables were most frequently reported types of variables. Convergent variables were also reported less frequently (13 %), whereas discriminant variables were not mentioned in any articles at this time. Comparisons between practice in the two time periods reveal an increase in reports of convergent (13 % in 1950–1960 to 42 % in 2000–2010) and discriminant evidence (from none in 1950–1960 to 37 % in 2000–2010) and a decrease in reports of the two types of criterion-related evidence over time (from 63 % in 1950–1960 to 32 % in 2000–2010 for predictive evidence and from 44 % in 1950–1960 to 16 % in 2000–2010 for concurrent evidence).

Discussion

The first aim of the current study was to document validation practice reported in validation articles published in JEP from 2000 to 2010 (Research Question 1). The results revealed that current practice reflects modern validity theory in several ways. In particular, most of the recent articles reported construct evidence, which reflects contemporary ideas about construct evidence being the whole of validity (Sireci 2009). In addition, most of the articles reported internal structure evidence. However, the results also revealed several ways in which validation practice did not reflect the current Standards (e.g., AERA et al. 1999) and modern validation theory (e.g., Zumbo 2007). For example, most articles referred to multiple types of validity and explained that validity was a property of a test. Furthermore, no articles reported evidence based on response processes or the consequences of testing.

The second aim of the current study was to compare practice across two time periods to see whether and how validation practice has changed over time (Research Question 2). Results revealed that the recent articles more regularly cited relevant experts, and used a wider variety of comparison variables. However, in several other categories practice was very similar in the two time periods. In particular, in both time periods the Standards (e.g., AERA et al. 1999) were not referenced, internal structure was reported as reliability evidence only (not as also informing validity), only one source of evidence was generally provided, and justification was rarely provided for comparison variables. From this, we can conclude that validation practice across the two time periods is similar in many ways despite the passing of 40–50 years and the publishing of four test standards (AERA et al. 1985, 1999; APA et al. 1966, 1974) during that time. Three major findings and their implications are discussed below.

Response Processes and Consequences of Testing

The results revealed that none of the articles published in 2000–2010 reported evidence of response processes or testing consequences. Given that validation occurs through a process of accumulation of evidence (AERA et al. 1999), it is not necessary for articles to include all sources of validity evidence. At the same time, however, the fact that no articles reported these two sources of evidence suggests that they may be being ignored. This has important implications.

As described above, evidence based on response processes refers to examinations of participants' responses and why they chose those responses (AERA et al. 1999). Response processes are helpful for understanding whether there are major individual differences in processes for answering questions, why this may have occurred, and how this may affect the responses (AERA et al. 1999). For example, evidence based on response processes can reveal differences in interpretations of test questions across different subgroups of participants. This is important for understanding whether the questions are accurately measuring the construct or whether some other type of variance is causing differences in scores (e.g., different meanings among different subcultures). It can also provide understanding of why this occurs, which can be used to create better instruments that more accurately capture the construct or knowledge under examination across different subgroups.

Examinations of response processes are also useful for developing definitions of a construct by revealing understanding about how it is interpreted by participants. This can also help ensure that participants are interpreting the questions as expected and, in turn, that their responses reflect the construct or knowledge that the researcher is attempting to examine. An example of this appeared in Gaderman et al.'s (2011) research. They examined the response processes of children as they answered the Satisfaction with Life Scale adapted for Children (Gaderman et al. 2011). The analyses revealed greater understanding of the construct by

showing that the children used strategies to answer the questions that were meaningful and theoretically consistent with the literature.

Also important is evidence based on test consequences. As described above, evidence based on test consequences aims to investigate the intended and unintended consequences of testing (AERA et al. 1999). As a broad statement, this type of evidence is important for considering construct irrelevant variance so that, as Hubley and Zumbo (2011) note, based on construct delineation and definition intended social and personal consequences and unintended social and personal side effects emerge. As summed up by Shepard (1997), "consequences are evaluated in terms of the intended construct meaning" (p. 8). With the domain of educational psychology in mind, as Hubley and Zumbo (2011) suggest, when the social consequences and side effects of using an educational psychology measure are not congruent with our societal values and goals regarding that particular psychological domain such insights in the validation process may be used to adjust constructs, theories, and aspects of the measurement process until the desired congruence between purposes, goals, values, and consequences is accomplished. For example, consider the construct of self-efficacy, which refers to individuals' beliefs about their capabilities in a prospective and specific context (Bandura 1982, 1993). When researchers measure this construct, a consequence may be the promotion of self-efficacy as a positive characteristic in educational and developmental contexts. This may be an intended consequence given that research has highlighted self-efficacy is important for positive outcomes among individuals (e.g., greater achievement among students; Caprara et al. 2006). However, it is important to note that self-efficacy is not necessarily positive at very high levels.

For example, Brenner et al. (2012) examined the development of self-efficacy for teaching among student teachers as they engaged in their practicum placement in schools during their teacher education program. Through their analyses, they described one student teacher who reported consistently high levels of self-efficacy for teaching despite receiving low ratings of effectiveness by her faculty advisor (i.e., the faculty member who assessed student teachers' progress). Moreover, Brenner et al. explained that the student teachers' high levels of self-efficacy may have prevented her from realistically assessing her abilities and putting in the necessary effort to improve her teaching skills. This example highlights that too much self-efficacy can, in fact, be a negative such that individuals may not feel the need to work on improving their own practice. When researchers assess constructs like self-efficacy, one consequence is that the construct becomes valued and participants may feel that they need to experience high levels of it. In most cases, this may be positive. Nevertheless, it is a consequence and should be considered.

When researchers do not consider evidence based on response processes and testing consequences, the implications potentially include a weaker understanding of the construct under examination and the promotion of certain outcomes among participants (that may or may not be positive). Clearly, greater efforts to include these types of evidence in validation practice are needed.

Reference to Standards

The second major finding to be discussed refers to references that articles made to the Standards (e.g., AERA et al. 1999). Considering that a proposal for test standards was first published in 1952 (APA, Committee on Test Standards) and that the first set of test standards were created in 1954 (APA), it is understandable that articles published before and shortly after this time (i.e., between 1950 and 1960) did not cite any test standards. However, all the articles published between 2000 and 2010—when five different versions of the test standards (AERA et al. 1985; APA et al. 1966, 1974; APA 1954) have been published—did not cite any version of the Standards. This result highlights a disconnect between validity theory and contemporary validation practice. If researchers do not cite the Standards in their publications, then it is questionable as to whether they are consulting them in their research, which leads to further concerns about the dissemination of validity theory.

There are two plausible reasons for why this finding occurred. First, it is possible that researchers are simply not familiar with the current Standards (AERA et al. 1999). If this is the case, there is a need for efforts that endeavor to increase the visibility of the current Standards and to educate researchers about how to use them in their practice. The second reason is that authors may be aware of the current Standards, but not feel it is necessary to make reference to them. For this possibility, journals like JEP and their editors may want to raise expectations about what constitutes accurate validation practice by requiring articles to accurately address modern validity theory in their validation practice. Clearly, more visibility is needed to raise general awareness of the Standards. However, expectations must also be raised to ensure that visibility is followed through with accurate practice.

Validity Characterization

The third key finding prevalent among the recent JEP validation articles was the presentation of several characteristics of validation practice that do not conform to modern validity theory. In particular, the view of validity as having multiple types and the conceptualization of validity as the property of a test highlight that validation practice has not kept pace with the changes in core aspects of validity theory. This is particularly concerning given that the unitary view of validity was first articulated in the 1950s by Loevinger (1957) and first presented in the 1985 Standards (AERA et al.). Furthermore, validity as adhering to inferences rather than a test was introduced in the 1985 Standards (Goodwin and Leech 2003). In other words, despite the fact that these aspects of validity theory had been published in the Standards (AERA et al. 1985, 1999) for 15–25 years when the recent articles were published, researchers are still using outdated practices.

As to why this occurred, it is possible that it relates to semantic differences (Cizek et al. 2008). Cizek and colleagues' (2008) suggest that despite the different nomenclature, *sources of evidence* and *types of validity* may actually have the same underlying meaning for researchers. However, it is also possible that this reflects confusion about the meaning and nature of validity. Given that both the current study and other studies (e.g., Cizek et al. 2008) have found the same misconceptions about types of validity and validity as a property of a test, this is an issue that warrants further consideration and investigation. Furthermore, validity theorists may want to carefully consider the language they use to explain validity theory in order to emphasize the importance of specific terminology, as well as the meaning underlying it. Clearer language and a better understanding of why certain language is used will likely help to increase the accessibility of the theory for researchers.

Another plausible reason for the prevalence of outdated practice is that researchers are unaware of how to provide evidence that conforms to modern validity theory. Instead of referring to the Standards (e.g., AERA et al. 1999), researchers may refer to existing examples of validation practice in the literature. When these examples utilize older theory and practices of validation, this creates a cycle of outdated practice informing more outdated practice. In order to remedy this, examples that are in line with contemporary validity theory are needed. As noted above, journals and journal editors may want to raise their expectations regarding the validation practice that is published in journals. By expecting researchers to report validation practice that is based in modern validity theory, more accurate examples will begin to appear in the literature. This will ideally disrupt the current cycle to create a shift towards better dissemination and practice.

Limitations

There are several limitations to the current study that must be discussed. First, the study examined 35 articles published in JEP between 1950 and 1960, and 2000 and 2010. Given the small sample, the generalizability of the results is limited to validation articles published in JEP around these two times. The second limitation is that we interpreted articles' provision of validity evidence as the position of the article, when it may in fact reflect an editorial position. Authors may have been discouraged from including too much information on validity to ensure the readability of the article, or they may have excluded it from their submissions for fear of being rejected from publication for being overly psychometric. The nature of our examination did not allow us to determine the reasons behind authors' reports of validation practice. For this reason, rather than referring to the authors we have referred to the articles, as they, not the author's perspectives were the data.

Conclusions

The current study has shown that despite great changes in validity theory and the Standards (e.g., AERA et al. 1985, 1999) over the past half century, current validation practice (presented in JEP articles) does not appear to wholly conform to modern validity theory. This is an issue that was raised by Messick (1988) two decades ago, and by Borsboom et al. (2004) more recently: "The concept that validity theorists are concerned with seems strangely divorced from the concept that working researchers have in mind when posing the question of validity" (p. 1061). The main conclusion from the current study, therefore, is that more must be done to help validation practice draw level with validity theory.

Although JEP is not a journal that focuses on psychometrics or psychological measurement, it is a journal published by the American Psychological Association (APA), a key author in all versions of the Standards (AERA et al. 1985; APA et al. 1966, 1974; APA 1954). Despite this, the Standards' recommendations do not appear to be required practice in JEP validation articles. Perhaps this reflects the different demographic of readers of JEP, or as suggested in the limitations above, the position of an editor. However, it certainly provides an interesting question regarding the practice of modern validity theory: If the APA, as a key organization in the creation of the Standards, does not require that authors follow the Standards in one of their own publications, then is it really surprising that there is such a disparity between theory and practice across the field? Considering the many experts who have questioned the gap between theory and practice (e.g., Borsboom et al. 2004; Hubley and Zumbo 1996; Messick 1988), this is certainly a question that needs to be addressed. As noted above, we recommend that journals raise their expectations, especially journals published by the AERA, APA, and NCME, so that validation articles accurately adhere to modern validity theory. This will not only provide examples for other researchers, but it will aid in the much-needed dissemination of the modern conceptualization of validity theory so that the gap between theory and practice is reduced.

Appendix

References of Articles Published Between 1950 and 1960

Ausubel, D. P., Schiff, H. M., & Zeleny, M. P. (1954). Validity of teachers' ratings of adolescents' adjustment and aspirations. *Journal of Educational Psychology, 45*, 394–406. doi:10.1037/h0056091.

Brown, W. F., & Holtzman, W. H. (1955). A study-attitudes questionnaire for predicting academic success. *Journal of Educational Psychology, 46*, 75–84. doi:10.1037/h0039970.

Chappell, T. L. (1955). Note on the validity of the army general classification test as a predictor of academic achievement. *Journal of Educational Psychology, 46*, 53–55. doi:10.1037/h0044316.

Doppelt, J. E., & Wesman, A. G. (1952). The differential aptitude tests as predictors of achievement test scores. *Journal of Educational Psychology, 43*, 210–217. doi:10.1037/h0060030.

French, J. W. (1958). Validation of new item types against four-year academic criteria. *Journal of Educational Psychology, 49*, 67–76. doi:10.1037/h0046064.

French, J. W. (1959). The relationship of home and school experiences to scores on achievement tests. *Journal of Educational Psychology, 50*, 75–82. doi:10.1037/h0047991.

Gage, N. L. (1957). Logical versus empirical scoring keys: The case of the MTAI. *Journal of Educational Psychology, 48*, 213–216. doi:10.1037/h0047795.

Holland, J. L. (1959). The prediction of college grades from the California psychological inventory and the scholastic aptitude test. *Journal of Educational Psychology, 50*, 135–142. doi:10.1037/h0041909.

Lodge, W. J. (1951). A validity study of personality questionnaires at the upper elementary grade level. *Journal of Educational Psychology, 42*, 21–30. doi:10.1037/h0061737.

Lorge, I., & Diamond, L. K. (1954). Validity of an objective examination for the placement of foreign students in English courses. *Journal of Educational Psychology, 45*, 208–214. doi:10.1037/h0053872.

Neidt, C. O., & Merrill, W. R. (1951). Relative effectiveness of two types of response to items of a scale on attitudes toward education. *Journal of Educational Psychology, 42*, 432–436. doi:10.1037/h0056066.

Papavassiliou, I. T. (1953). The validity of the goodenough draw-A-man test in Greece. *Journal of Educational Psychology, 44*, 244–248. doi:10.1037/h0057111.

Schultz, D. G. (1954). Item validity and response change under two different testing conditions. *Journal of Educational Psychology, 45*, 36–43. doi:10.1037/h0059845.

Scott, O., & Brinkley, S. G. (1960). Attitude changes of student teachers and the validity of the Minnesota Teacher Attitude Inventory. *Journal of Educational Psychology, 51*, 76–81. doi:10.1037/h0040593.

Uhlinger, C. A., & Stephens, M. A. (1960). Relation of achievement motivation to academic achievement in students of superior ability. *Journal of Educational Psychology, 51*, 259–266. doi:10.1037/h0041083.

Woodcock, R. W. (1958). An experimental prognostic test for remedial readers. *Journal of Educational Psychology, 49*, 23–27. doi:10.1037/h0042526.

References of Articles Published Between 2000 and 2010

Bong, M. (2009). Age-related differences in achievement goal differentiation. *Journal of Educational Psychology, 101*, 879–896. doi:10.1037/a0015945.

Branum-Martin, L., Mehta, P. D., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M., & Francis, D. J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology, 98*, 170–181. doi:10.1037/0022-0663.98.1.170.

Brockway, J. H., Carlson, K. A., Jones, S. K., & Bryant, F. B. (2002). Development and validation of a scale for measuring cynical attitudes toward college. *Journal of Educational Psychology, 94*, 210–224. doi:10.1037/0022-0663.94.1.210.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*, 891–901. doi:10.1037/0022-0663.98.4.891; 10.1037/0022-0663.98.4.891.supp (Supplemental).

Chowning, K., & Campbell, N. J. (2009). Development and validation of a measure of academic entitlement: Individual differences in students' externalized responsibility and entitled expectations. *Journal of Educational Psychology, 101*, 982–997. doi:10.1037/a0016351.

Craven, R. G., Marsh, H. W., Debus, R. L., & Jayasinghe, U. (2001). Diffusion effects: Control group contamination threats to the validity of teacher-administered interventions. *Journal of Educational Psychology, 93*, 639–645. doi:10.1037/0022-0663.93.3.639.

d'Ailly, H. (2003). Children's autonomy and perceived control in learning: A model of motivation and achievement in Taiwan. *Journal of Educational Psychology, 95*, 84–96. doi:10.1037/0022-0663.95.1.84.

Edwards, W. R., & Schleicher, D. J. (2004). On selecting psychology graduate students: Validity evidence for a test of tacit knowledge. *Journal of Educational Psychology, 96*, 592–602. doi:10.1037/0022-0663.96.3.592.

Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100*, 613–628. doi:10.1037/0022-0663.100.3.613.

Gathercole, S. E., & Pickering, S. J. (2000). Assessment of working memory in six- and seven-year-old children. *Journal of Educational Psychology, 92*, 377–390. doi:10.1037/0022-0663.92.2.377.

Greene, J. A., Torney-Purta, J., & Azevedo, R. (2010). Empirical evidence regarding relations among a model of epistemic and ontological cognition, academic performance, and educational level. *Journal of Educational Psychology, 102*, 234–255. doi:10.1037/a0017998.

Grigorenko, E. L., Jarvin, L., Diffley, R., Goodyear, J., Shanahan, E. J., & Sternberg, R. J. (2009). Are SSATS and GPA enough? A theory-based approach to predicting academic success in secondary school. *Journal of Educational Psychology, 101*, 964–981. doi:10.1037/a0015906.

Janosz, M., Le Blanc, M., Boulerice, B., & Tremblay, R. E. (2000). Predicting different types of school dropouts: A typological approach with two longitudinal samples. *Journal of Educational Psychology, 92*, 171–190. doi:10.1037/0022-0663.92.1.171.

Kardash, C. M., & Wallace, M. L. (2001). The perceptions of science classes survey: What undergraduate science reform efforts really need to address. *Journal of Educational Psychology, 93*, 199–210. doi:10.1037/0022-0663.93.1.199.

Legault, L., Green-Demers, I., & Pelletier, L. (2006). Why do high school students lack motivation in the classroom? Toward an understanding of academic amotivation and the role of social support. *Journal of Educational Psychology, 98*, 567–582. doi:10.1037/0022-0663.98.3.567.

Mokhtari, K., & Reichard, C. A. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology, 94*, 249–259. doi:10.1037/0022-0663.94.2.249.

Naglieri, J. A., & Rojahn, J. (2004). Construct validity of the PASS theory and CAS: Correlations with achievement. *Journal of Educational Psychology, 96*, 174–181. doi:10.1037/0022-0663.96.1.174.

Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology, 98*, 598–616. doi:10.1037/0022-0663.98.3.598.

Watkins, M. W., & Coffey, D. Y. (2004). Reading motivation: Multidimensional and indeterminate. *Journal of Educational Psychology, 96*, 110–118. doi:10.1037/0022-0663.96.1.110.

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

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

American Psychological Association, Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist, 7*, 461–465.

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.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*, 122–147.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*, 117–148.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071. doi:10.1037/0033-295X.111.4.1061.

Brenner, C. A., Perry, N. E., & Collie, R. J. (2012, September). Student teachers' developing practices that promote self-regulated learning: Linking efficacy and utility beliefs to effectiveness. In N. Perry & B. Kramarski (Co-chairs), *Metacognition and self-regulation in developing professionals*. Symposium presented at the biennial meeting of the European Association for Research on Learning and Instruction, Milan, Italy.

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

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *Journal of School Psychology, 44*, 473–490. doi:10.1016/j.jsp.2006.09.001.

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. doi:10.1177/0013164407310130.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, L. J. (1949). *Essentials of psychological testing*. New York: Harper.

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

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the satisfaction with life scale adapted for children: A focus on cognitive processes. *Social Indicators Research, 100*, 37–60. doi:10.1007/s11205-010-9603-x.

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist, 27*, 197–222.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*(3), 181–191.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123*, 207.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research: An International Interdisciplinary Journal for Quality of Life Measurement, 103*, 219–230. doi: http://dx.doi.org/10.1007/s11205-011-9843-4.

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*, 736–753. doi:10.1177/0013164498058005002.

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

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale: Erlbaum.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. doi:10.1037/0003-066X.50.9.741.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16*, 296.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 19, pp. 405–450). Washington, DC: AERA.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5–8, 13, 24.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83–117.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte: Information Age Publishing.

Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook*. Lincoln: Buros Institute of Mental Measurements.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Psychometrics, Vol. 26, pp. 45–79). Amsterdam: Elsevier Science.

Chapter 8
A Review of Validity Evidence Presented in the Journal of Sport and Exercise Psychology (2002–2012): Misconceptions and Recommendations for Validation Research

Katie E. Gunnell, Benjamin J.I. Schellenberg, Philip M. Wilson, Peter R.E. Crocker, Diane E. Mack, and Bruno D. Zumbo

Introduction

Measurement lies at the heart of any quantitative research design in sport and exercise psychology. The inferences that researchers make from measurement must therefore be based on the assumption that the instrument used in assessment produces scores that are both valid and reliable. Score validity is therefore a foundational aspect for creating, developing and using instruments (American Educational Research Association [AERA] et al. 1999). Validity is defined as "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (AERA et al. 1999, p. 9). Without score validity, any inference made from an instrument is meaningless (Hubley and Zumbo 1996). In sport and exercise psychology, researchers rely heavily on the use of instruments designed to tap various psychological constructs (e.g., anxiety, well-being, motivation). It is essential that scores derived from these instruments demonstrate

K.E. Gunnell (✉) • P.R.E. Crocker
School of Kinesiology, The University of British Columbia, 210 War Memorial Gym, 6081 University Boulevard, Vancouver, BC V6T 1Z1, Canada
e-mail: kgunnell@alumni.ubc.ca

B.J.I. Schellenberg
Department of Psychology, The University of Manitoba, Winnipeg, MB R3T 2N2, Canada

P.M. Wilson • D.E. Mack
Behavioural Health Sciences Research Lab, Department of Kinesiology, Brock University, Niagara Region, 500 Glenridge Ave., St. Catharines, ON L2S 3A1, Canada

B.D. Zumbo, Ph.D. (✉)
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_8,

validity evidence such that inferences and conclusions forwarded remain trustworthy. Unfortunately, validation practices by researchers in sport and exercise psychology have been characterized as outdated and in need of improvement (Zhu 2012). For example, it has been speculated that researchers are using outdated language when referring to score validity, confusing validity terms and incorrectly labelling evidence of validity (e.g., convergent evidence is reported as concurrent evidence; Zhu 1998), and/or altogether neglecting validity theory when conducting validation studies (Gunnell et al., Chap. 10, this volume).

While attempts have been made to encourage stronger validation endeavours (e.g., Hagger and Chatzisarantis 2009; Zhu 2012), there has been little attempt to quantify exactly how validation is being conducted in the sport and exercise psychology literature. Furthermore, review articles that focus on the concept of validity are sometimes outdated and lack information pertaining to current advances in validity theory (e.g., Hagger and Chatzisarantis 2009). Therefore, the purpose of this investigation is to examine how researchers who have published in a premier journal in the field of sport and exercise psychology have encapsulated (or omitted) aspects of contemporary validity theory in their validation investigations. This purpose will be achieved using a systematic review combined with a narrative commentary. This paper will first briefly outline validity theory and different validation frameworks. Following this brief overview, the methods, results, and discussion specific to the systematic review and narrative commentary will be presented. Finally, recommendations on how to improve the current state of validation research in sport and exercise psychology will be outlined.

Validity Theory

Validity theory and the process of score validation are as old as testing and measurement itself (for an accessible review of the history of validity theory, see Sireci 2009). Validity theory is complex, and not all experts agree on every aspect outlined. There is, however, a general consensus that (a) validity refers to test scores and interpretations. That is, an instrument is neither valid nor invalid. Validity is a property of the scores obtained from an instrument and what we attempt to validate are the scores from the instrument, not the instrument itself per se (Sireci 2009) and (b) validation is an ongoing process (Zumbo 2007) whereby it is fallacious to assume an instrument is valid across all contexts and/or samples. The following section is designed to provide a brief overview of various validation frameworks from which researchers can base their validation efforts.

The *Standards*

The Standards for Educational and Psychological Testing (*The Standards*; AERA et al. 1999) is a validity theory and validation manual developed through

collaborations between the AERA, APA and NCME, and committees made up of prominent psychometritions and validity theorists such as Cronbach, Meehl, and Messick. The *Standards* can be viewed (and used) as a general guideline based on sound validity theory for evaluating score validity.

The current edition of the *Standards* lists five sources of validity evidence based on (a) content, (b) internal structure, (c) relations to other variables, (d) response processes, and (e) consequences. The *Standards* encourage researchers to describe "validity evidence" and denounce the nomenclature of validity 'types'. This shift from validity 'type' to 'evidence' was created in an attempt to help researchers realize that validity is a unified concept and it is inappropriate to select certain 'types' of validity while ignoring others. As a unified concept, all validity evidence bears on score validity whereby different sources of evidence can illuminate sources of invalidity. Validity *evidence based on content* refers to the content of the construct, including wording, themes, item format, and guidelines for administering the instrument (e.g., expert views of item content; Dunn et al. 1999). Validity *evidence based on internal structure* refers to the degree to which instrument items conform to the construct from which the scores are interpreted (e.g., factor analysis, item interrelationships). Validity *evidence based on relations to other variables* is defined as the analysis of relationships among test scores to variables *external* to the instrument (e.g., criterion predictive, convergent). Validity *evidence based on response processes* concerns the degree to which the scores are actually measuring the construct versus the responses actually engaged in by the participant (e.g., think aloud protocols to determine how participants are interpreting and answering the items). Finally, validity *evidence based on consequences* of testing concerns the intended and unintended consequences of test score use that informs validity decisions (e.g., if the instrument produces biased scores, it may have unintended consequences for the population it is biased against).

Education and psychology are considered to be the parent academic disciplines that gave rise to the field to sport and exercise psychology (Zhu 2012). As a consequence, our knowledge (albeit outdated) relies on theories and information developed in education and psychology. Therefore, the current paper will utilize the *Standards* to address the main study purpose. As noted above, the *Standards* can be viewed as an overarching guiding framework for conducting score validation investigations, yet alternative validity frameworks have been forwarded by psychometricians such as Messick (1995), Kane (2001, 2013), Zumbo (2007), and Borsboom et al. (2004). A very brief overview of other validity frameworks and theories will be outlined next. With the exception of work by Borsboom and colleagues (2004) on validity, Messick, Kane, and Zumbo offer new insights to understanding score validity that are largely complimentary to the *Standards*.

Messick's Progressive Matrix

Messick (1995) advocated a progressive matrix of validity based on construct validity. At the centre of every validation effort lies construct validity as an

integrating facet that brings together validity concepts to form a unitary conception of validity. Messick outlines six aspects of construct validity including: (1) content, (2) substantive, (3) structural, (4) generalizability, (5) external, and (6) consequential. The content aspect is similar to that described by the *Standards* and includes content relevance and representativeness. The substantive aspect describes the theoretical rationales used to create and evaluate instruments, and also the processes actually engaged in by participants (similar to the *Standards* response processes). The structural aspect refers to the scoring structure, and the generalizability aspect refers to the examination of generalizability of score properties across populations, settings, or tasks and the boundaries of score meaning. The external aspect refers to the relationships between variables external to the instrument (e.g., convergent and discriminant correlations; similar to the *Standards* relationships to other variables). Finally, Messick, a main proponent of consequences as validity evidence, argues that validity evidence that evaluates the intended and unintended consequences be assessed. Together, the six aspects of validity evidence constitute a means of examining central issues related to the unified concept of validity (Messick 1995). A final contribution Messick made to validity theory was the progressive matrix. Messick (1995) contends that validity is comprised of two interrelated facets. The first facet concerns the source of justification of the testing (evidential or consequential) and the second facets concerns the outcome of testing (interpretation or use). For more information on the progressive matrix of validity, please see Messisk or Hubley and Zumbo (1996).

Kane's Argument Based Approach

Kane explicitly outlines a 'validity argument' wherein an argument based on inferences and uses from instruments is made (Kane 2013). The argument based approach involves different steps for assembling the validity argument (see Kane 1992, 2001, 2013). All the available evidence is gathered and each validity argument is only as strong as the weakest link. Kane also highlights the notion of weak and strong forms of validity. In the weak form, correlations with other variables are considered as score validity. In the strong form, theoretical rationales are made explicit and then deliberately challenged. An important part of the argument is testing competing sources of evidence or interpretations of the scores (Sireci 2009). Finally, Kane acknowledges that a validity argument can never be absolute, and researchers must therefore gather as much evidence as possible to make a meaningful argument based on the kind of claim being made (Kane 2013). That is, a more ambitious claim requires more strong evidence (Kane 2013).

Zumbo's Draper-Lindley-deFinetti (DLD) Framework

Within the Draper-Lindley-deFinetti (DLD) framework, Zumbo (2007) calls attention to the assumptions that must be tested to investigate score validity. At the

forefront is sample homogeneity. Discussed are the implications of sampled and unsampled respondents and whether they are exchangeable in a target population. In so doing, the limits or bounds of inferences derived from instruments are specifically examined. Zumbo lists four forms of inference: (a) initial calibrative inference in which the inference cannot extend beyond the sample from which the data were gathered, (b) specific sampling inference where inferences can be extended to the specific sample from which the data were gathered, (c) specific domain inference where inferences can be made about what is being measured, and (d) general measurement inference that permits comparisons across measures and different samples. In essence, Zumbo's DLD framework brings attention to the context, bounds of inferences, and sample heterogeneity.

Borsboom, Mellenbergh, and Heerden's Construct Validity

Finally, Borsboom advances a theory of validity that stands in direct contrast to construct validity and a unified theory of validity (Borsboom et al. 2004). Within the theory espoused by Borsboom and colleagues, validity is conceptualized as a property of the instrument, and validity lets a researcher know if the instrument is sensitive to changes in the variable it is predicting (Borsboom et al. 2009). That is, a test is valid if an attribute exists and that attribute causally predicts variation in an outcome (Borsboom et al. 2004).

Study Purpose and Research Questions

Although the most recent edition of the *Standards* was released over 10 years ago, many researchers in sport and exercise psychology rely on outdated historical definitions and antiquated conceptualizations of validity (Zhu 2012). A cursory glance at research in sport and exercise psychology would reveal that many researchers still conceptualize validity as a property of the instrument (e.g., Gucciardi 2011; Hagger and Chatzisarantis 2009). This type of statement assumes that validity is a stable property of the instrument. This conceptualization is in stark contrast to the *Standards* that describe validity as a dynamic property of the scores derived from the instrument for any given sample under study. Zhu (2012) also critiqued researchers in sport and exercise psychology for relying too heavily on evidence of internal structure, a trend noted by Gunnell and colleagues (Chap. 10, this volume). Furthermore, the current version of the *Standards* (AERA et al. 1999) clearly emphasizes the distinction between 'types' of validity and types of evidence for score validity, denouncing the use of "traditional nomenclature (i.e., the use of the terms content validity or predictive validity" [AERA et al., p. 11]). Rather, the use of the word 'evidence' (e.g., convergent evidence) should be used as part of a strong validity argument (AERA et al. 1999; Kane 2001) in an effort to avoid the issue of validity types. Yet many researchers in sport and exercise psychology still

use terms such as ‘content validity’ and ‘convergent validity’, implying that different types of validity exist in a manner commensurate with the ‘holy trinity’ that is now considered to be an outdated approach to understanding validity (Zhu). Overall, it seems reasonable to suggest that research outlining the state of validity theory and validation practices over a wide time span (i.e., 10 years vs. the 1 year investigation conducted by Zhu), and validation efforts published in a high impact journal in the field would be a welcome addition to the sport and exercise psychology literature.

The purpose of this investigation is to examine how researchers in sport and exercise psychology, who publish their validation research in *JSEP* report validity information. More specifically, because previous research has documented concerns with regard to how researchers reported validity evidence (Gunnell et al., Chap. 10, this volume; Zhu 2012), we were interested in conducting a more comprehensive and rigorous examination by (a) only including investigations that indicated score validation was a main study purpose, (b) including all investigations over the past 10 years, and (c) examining investigations published in the leading journal devoted to research in sport and exercise psychology (i.e., *JSEP*, Impact Factor = 2.658). Furthermore, we wanted to determine if validation studies published in this leading journal adhered to validity theory and validation protocols advanced by governing bodies such as the *Standards*. Specific research questions were as follows:

1. What perspectives or validity framework were researchers basing their validation investigations on?
2. What sources of validity evidence were researchers reporting in their validation investigations?
3. How frequently were researchers investigating different aspects of score validity?
4. Were there any misconceptions about validity theory and validation?

Methods

Sampling

A computerized search was conducted to locate all articles published in *JSEP* between 2002 and 2012 (journal volumes 24–34). In total, 405 journal articles were identified and screened for inclusion/exclusion criteria.

Inclusion/Exclusion Criteria

Journal articles were included if their title, abstract, or keywords contained any of the following words: validity, validation, valid. The first author screened all journal articles for inclusion/exclusion criteria. Of the original 405 articles screened, 58 articles met the inclusion criteria. Articles that met the inclusion criteria were then coded independently by two authors. Eight articles were removed because they did not directly evaluate validity evidence (e.g., the abstract contained the word validity to describe an experimental cue).

Data Collection

A standardized coding scheme was used in order to reduce ambiguity between coders. The first author and coder had training in validity theory, research methods and statistics at the doctoral level, and had a publication record in the area of validation. The second coder was a doctoral student who had completed coursework relevant to validation including research methods, statistics, and psychological testing. The first author trained the second coder on aspects related to validity theory and validation. The second coder familiarized himself with validity theory through select readings (e.g., AERA et al. 1999; Goodwin 2002; Cizek et al. 2008). The first author coded all journal articles. Prior to the second author coding all articles, two articles with high complexity were selected to code. The primary and secondary coder independently coded these two articles and then discussed the results of the coding to determine the coding scheme's clarity, discuss ambiguity, and to familiarize the second coder with the coding process. Once the coding manual was deemed to be clear, the second coder independently coded all journal articles. Percent agreement between the first and second coder was above 78 %. When a difference in coding occurred, the two coders discussed the difference and resolved the discrepancies prior to data analysis. All discrepant coding was resolved without the need for a third evaluation.

Coding Scheme

A standardized coding scheme was created using definitions based on the *Standards* (AERA et al. 1999) and Goodwin's (2002) review of the *Standards*.[1] Various aspects of journal articles meeting inclusion criteria were coded such as: (a) article type (e.g., research, review, position statement, editorial, psychometric and unsure/not clear), (b) study type (e.g., qualitative, quantitative, mixed methods,

[1] The coding scheme is available from the first author upon request.

unsure/not clear), (c) data analysis (e.g., meta-analysis, systematic review, statistical analysis, unsure/not clear), (d) if the authors reported validity evidence (yes/no), and (e) if the article contained multiple studies and/or samples (yes/no). Validity information coded was based on previous research (Cizek et al. 2008) and the *Standards* (AERA et al. 1999) and included (a) if the authors reported reliability evidence as validity evidence (yes/no), and (b) the validity perspective used (unitary perspective, the *Standards*, Messick, other, and unsure/not clear). Next, the evidence of score validity presented in each manuscript was coded based on the sources outlined by the *Standards* (evidence based on content, internal structure, relations to other variables, response processes, and/or consequences). Each of these five categories was further subdivided to provide a clearer overview of the information within each study. Evidence based on content was coded as: (a) content, (b) face, (c) other, (d) unsure/not clear. Evidence based on internal structure was coded as: (a) factor analysis, (b) item interrelationships, (c) invariance, (d) other, (e) unsure/not clear. Relations to other variables was coded as: (a) convergent, (b) divergent, (c) discriminant, (d) criterion predictive, (e) criterion concurrent, (f) criterion group differences, (g) validity generalizability, (h) construct validity,[2] (i) other and (j) unsure/not clear. Evidence based on response processes was coded as: (a) analysis of individual responses via interview with test takers, (b) monitoring changes or development of responses, (c) process studies, similarities or differences in responses given by members of distinct groups, judges, or observers or interviewers collect record and interpret data. Finally, evidence based on consequences was coded as: (a) benefits are tested, (b) negative consequences are tested, (c) other, (d) unsure/not clear. It is important to note that validity information was coded based on what the authors of the investigation reported. For example, if authors reported that they provided evidence of "concurrent validity" but the results indicated that it was evidence of convergence, the validity information was coded as concurrent validity and a note was made indicating that the authors mislabelled the type of evidence.

Results

Coded Studies

Out of the 50 studies retained for coding, 37 (74 %) reported evidence from more than one study or sample. One investigation was qualitative, 34 (68 %) were quantitative and 15 (30 %) used mixed methods. The majority of the investigations ($n = 44$, 88 %) were coded as having validation as the primary purpose while six

[2] Some authors refer to all validity evidence as construct validity (i.e., a unified perspective). Construct validity was only coded as 'construct validity' under relations to other variables if the authors described it as a type of evidence bearing on the relationship with other variables.

Table 8.1 Validity perspectives

Validity perspective	Number of studies
Unitary perspective	7
Standards	1
Messick	4
Unsure/not clear	35
Other	3

investigations were coded as 'research', yet reported evidence of validity as complementary analysis.

What Perspective or Validity Framework Were Researchers Basing Their Validation Investigations On?

The majority of the coded studies presented no evidentiary basis in any validity theory framework (see Table 8.1). The most common validity theorist cited was Messick, followed by Cronbach. Only one investigation cited the *Standards*.

What Sources of Validity Evidence Were Researchers Reporting in Their Validation Investigations?

The most frequently reported source of validity evidence was internal structure, followed by relations to other variables, evidence of content, and response processes. None of the coded studies in any publication reported evidence of test consequences (see Table 8.2).

How Frequently Were Researchers Investigating Different Aspects of Score Validity?

When researchers were investigating internal structure, all examined factor structure through exploratory factor analysis, principal components analysis, or confirmatory factor analysis. Item interrelationships and invariance were examined less frequently than factor structure (see Table 8.2). One investigation examined a simplex structure and referred to it as construct validity. When examining evidence of content, the majority of researchers described their evidence as content with the term face validity used less often. When researchers were examining relations to other variables, the most frequently reported aspects were discriminant, convergent, concurrent, predictive, construct, divergent, group differences, nomological networks, or other, with the remaining studies representing the category labelled as unsure/not clear (see Table 8.2). Only one investigation examined response processes, with the investigators using think aloud protocols to examine how participants were responding to test items.

Table 8.2 Sources of validity evidence

Source of evidence	Number of studies
Content	18
Content	16
Face	4
Internal structure	46
Factor analysis	46
Item interrelationships	2
Invariance	16
Other: simplex pattern	1
Relations to other variables	39
Convergent	17
Divergent	1
Criterion-predictive	8
Criterion-concurrent	12
Criterion-group differences	2
Generalizations	0
Discriminant	20
Nomological network	4
Construct validity	8
Other	11
Unsure/not clear	2
Response processes	1
Consequences	0

Were There Any Misconceptions About Validity Theory and Validation?

Based on the investigation by Gunnell et al. (Chap. 10, this volume), we suspected that there would be discrepancies between how the *Standards* conceptualizes validity theory and validation compared with how applied researchers operationalize validation in sport and exercise psychology research. Specifically, coders recorded instances in which authors described validity as (1) a property of the test (vs. score or inference), (2) referred to 'types' of validity (vs. evidence of score validity), and (3) how authors examined evidence of convergent, discriminant, predictive and concurrence. Based on the coder notes, three prominent misconceptions were evident in the coded studies from *JSEP*.

Misconception 1: Validity as a Property of the Instrument

Authors of the investigations coded made references to validity being a property of the instrument under scrutiny. For example, statements such as "the SEQ is proposed as a valid measure of precompetitive emotion…" (Jones et al. 2005, p. 407), or "these findings lend support for the validity of the CMTI as a valid measure among adolescent cricketers…" (Gucciardi 2011, p. 370). Very few investigators consistently referred to validity as a property of a score, inference or response (e.g., Lonsdale et al. 2008). Many more used a mixture of language,

referring to validity as both a property of the instrument and as a property of the scores (e.g., Bartholomew et al. 2010; Sebire et al. 2008).

Misconception 2: Validity as a 'Type'

Most of the investigators used a mixture of language to describe validity; investigators referred to both validity 'types' and evidence of validity. For example, Bartholomew and colleagues (2010) referred to types of "factorial validity" (p. 193) and described validity as evidence showing "...support for the factor structure of the CCBS" (p. 209). A large portion of investigators exclusively referred to validity types such as "factorial validity" (Markland and Tobin 2004, p. 193), "content validity", "concurrent validity" (Williams and Cumming 2011, p. 419 and p. 432, respectively). Very few researchers exclusively used language that is advocated by the *Standards* (e.g., Myers et al. 2012).

Misconception 3: Incorrect Labels

Many investigators stated that they were examining 'criterion concurrent validity' by correlating scores from one instrument with scores of another, theoretically related instrument (e.g., Jones et al. 2005; Williams and Cumming 2011). This would be categorized by the *Standards* as an examination of convergent evidence unless the other instrument is a *criterion* measure that assesses the *same* construct. A few investigators did examine concurrent evidence correctly (e.g., Freeman et al. 2011). A separate issue noted while coding articles was that many authors described 'discriminant validity' after obtaining small-to-moderate correlations between instrument subscales (e.g., Boardley and Kavussanu 2007; Lonsdale et al. 2008). According to the current edition of the *Standards*, discriminant evidence is found using correlations between two instruments that are purported to measure different constructs, *not* subscales within one instrument.[3] Similarly, a few researchers noted convergent evidence when results of their factor analysis indicated that items loaded onto their respective latent factors (e.g., Williams et al. 2012). Convergent evidence is established using two *different* instruments that should theoretically be related (Campbell and Fiske 1959). Finally, many investigators examined predictive evidence using another criterion measure but did not include a time lag between assessments.

[3] We recognize that different researchers have different conceptualizations of discriminant and convergent evidence (e.g., Brown 2006; Kline 2010). If researchers are going to use alternative ways of examining convergent and discriminant evidence, they should ensure they cite relevant sources to support their analysis.

Discussion

Using a systematic review and narrative commentary, the purpose of this paper was to examine how researchers conducting validation studies published in *JSEP* have presented validity evidence. Results of the systematic review revealed that although a few researchers cited a validity theory framework, in general, researchers within the coded studies did not base their efforts on validation guidelines such as those advocated by the *Standards* (AERA et al. 1999), Messick (1995), or Kane (2001). Results also revealed that evidence of internal structure, relationships with other variables and content evidence were the most common sources of validity evidence presented in *JSEP* publications between 2002 and 2012. Validity evidence based on response processes and consequences has been largely omitted from validation efforts. Finally, results of this systematic review corroborate Zhu's (2012) contention that researchers in sport and exercise psychology conducting validation investigations published in *JSEP* are using outdated terminology, or mislabelling sources of evidence.

Validity Perspectives

Sireci (2009) has suggested that many validation efforts remain unsystematic. This systematic review substantiated Sireci's claims and illuminated a troubling trend whereby researchers conducting validation investigations are generally not basing their work on validity theory or any validation framework. This finding is consistent with one other investigation in exercise psychology (Gunnell et al., Chap. 10, this volume) and a systematic review from psychology (Cizek et al. 2008). Investigations devoted to examining the psychometric properties of an instrument's scores should explicitly outline the theoretical framework of validity utilized such that their use of validation techniques is justified. It is insufficient to use a hodgepodge of validity conceptions to piece together evidence of score validity without any clear framework. Such a fragmented approach to the analysis of score validity may actually confuse readers and fellow researchers and cloud appraisals of score validity for a specific instrument or population. By explicitly specifying a validity framework or validation process, and using standardized definitions (Zhu 2012), researchers can determine what evidence still needs to be gathered (e.g., response processes), or where evidence is insufficient (e.g., item interrelationships).

Validity Evidence, a Unified Concept

Zhu (2012) critiqued researchers in sport and exercise psychology for conducting one-shot validation studies. Coded studies in this investigation should be praised for

using multiple samples combined with multiple sources of validity evidence in their investigations. It is promising to know that researchers conducting validation studies are, with only a few exceptions, using different statistical techniques and providing more than one source of score validity evidence. Many coded studies reported up to three sources of validity evidence, with one studying reporting four (see Morton et al. 2011)!

Notwithstanding the importance of this finding, current validity theory is regarded as a unified concept (AERA et al. 1999; Messick 1995) and evidence based on response processes and test consequences have important meaning for understanding score validity. Only one investigation provided evidence based on response processes and no investigation provided evidence of consequences. Evidence based on response processes has the potential to inform the degree of fit between the theoretical construct and the detailed nature of the response the participant actually engages in when responding (AERA et al. 1999). For example, examining response processes could shed light on the degree to which social desirability is affecting the participants' responses to test items. Morton and colleagues (2011) employed a prospective think-aloud protocol to understand how participants interpret and respond to items. Through this process, the researchers were able to determine which items needed to be reworked, eliminated, or could be used in their original form. Using response processes can also inform researchers on how different sub-groups (e.g., sexes, ages, cultures) respond to test items (AERA et al. 1999).

Evidence based on test consequences has received attention over the last few decades (cf. Messick 1995). When instruments are developed, they are typically developed with the intention of being used in a particular way, be it for research or applied mental skills consulting for sport or exercise. When someone proposes to use an instrument in a particular way, such as for a screening tool to determine levels of trait anxiety, the use of this instrument must be justified by showing that the positive consequences outweigh the negative consequences (Kane 2001). For example, Myers and colleagues (2012) argued for the need of a referee efficacy scale because "efficacious referees should be more accurate in their decisions, more effective in their performance, more committed to their profession, have more respect from coaches, administrators, ... than less efficacious referees" (p. 738). It is entirely conceivable that the referee self-efficacy scale could be used by league officials when selecting referees to officiate important games (e.g., Super Bowl, Champions League final, etc.). Building on this example, it would be paramount to ensure validity evidence based on test consequences since the results of this test could lead to selection or de-selection of referees for important games.

Misconception 1: Validity of Score vs. Instruments

The distinction between an instrument being valid and a score being valid has significant implications because it emphasizes that the inferences made about

scores are bound by place, time and the use of the scores (Zumbo 2007). An overwhelming number of investigations did not adhere to the guidelines outlined by the current version of the *Standards* (AERA et al. 1999) in terms of language used to describe construct validity and score validation research. Few investigations reporting validity evidence accurately described it as a characteristic of the score/inference, and many more used a mix of language. It is important for researchers in sport and exercise psychology to reframe their language when discussing score validity such that outdated conceptions are not propagated. Implying that validity is a property of an instrument implies that future researchers, especially naïve researchers, can use the instrument beyond the context from which score validity evidence is available without considering the implications toward score meaning. For example, if researchers assumed a sport motivation scale developed using competitive athletes is 'valid', another researcher may use this 'valid' sport motivation scale to study recreational athletes without realizing the new sample may represent a source of invalidity that interferes with score interpretations. Thompson (1992) eloquently described the problem: "this is not just an issue of sloppy speaking- the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcomes, sloppy thinking and sloppy practice" (p. 436).

Misconception 2: Validity 'Type' vs. the Unified View

Recognizing the unified conceptualization of validity and moving away from old nomenclature, validity is no longer comprised of different 'types' but rather, various *sources of evidence* bear on score validity. The results of the current systematic review are consistent with previous research (Cizek et al. 2008; Zhu 2012) and provide evidence that researchers in sport and exercise psychology are still conceptualizing this portion of the measurement process incorrectly by referring to validity 'types'. It is not surprising that this misconception persists, given that a recent review pertaining to score validity published in this domain stated that "researchers in sport and exercise should be mindful of five types of validity…" (Hagger and Chatzisarantis 2009, p. 512). Validity theorists have moved away from the characterization of validity as distinct types because it creates problems with researchers forming check-list approaches (Hubley and Zumbo 1996), or interpreting it to mean that once one 'type' of validity was ascertained, alternative 'types' are not needed It is important that researchers maintain a standardized language when reporting evidence of score validity because it will prevent misconceptions that validity types represent distinct components of validity. Moreover, abandoning the validity 'type' language will emphasize the unified view of validity and keep researchers on the path to examining all aspects of score validity.

Misconception 3: Incorrectly Labelling Evidence of Validity

Based on previous contentions (Zhu 2012), it is not surprising to find that researchers within the coded investigations used various labels to assess the same aspect of validity evidence (e.g., using convergent and concurrent validity synonymously). It was, however, promising to note that many investigators used multitrait-mulitmethod (MTMM; Campbell and Fiske 1959) methods to assess discriminant and convergent evidence. Campbell and Fiske (1959) were the theorists who coined the term convergent and discriminant validity. In its original conception, evidence of convergence and discrimination was evaluated through correlations between scores of *different* instruments. That is, "convergence between independent measures of the same trait and discrimination between measures of different traits" (Campbell and Fiske, p. 104). The current iteration of the *Standards* still advocates that discriminant evidence and evidence of convergence be ascertained through correlations on test scores and scores from *other* instruments, not other subscales. While we do acknowledge that different authors describe convergent and discriminate evidence using different definitions than those advocated by psychometricians (e.g., Brown 2006), we want to stress the importance of conceptualization, and labelling sources of validity evidence that have been supported by validity theory and theorists. Using standardized terminology will help steer sport and exercise psychology researchers into the modern era and likely result in more rigorous validation studies. At the very least, authors should provide a reference to support their conceptualization of how validity was assessed.

Future Directions and Recommendations

While select results from the analyses reported in this systematic review are encouraging, it is evident that future research in sport and exercise psychology could benefit from the adoption of modern validity theory and validation practices. We can only speculate on why research in sport and exercise psychology lies far behind our parent fields (psychology and education). Some theorists acknowledge that modern approaches to psychometrics are challenging (Cizek et al. 2008; Shepard 1993), or, it is possible that researchers are simply not aware that a rich history for validity theory exists. In either case, there are solutions and recommendations that can be forwarded to provide impetus for sport and exercise psychology researchers to adopt an approach to psychometric investigations that is more aligned with 'state-of-the-art' recommendations. In this final section, we offer recommendations designed to advance the reporting of validity evidence in line with current approaches to validity theory for sport and exercise psychology researchers.

Further Education

Twenty years ago, Schutz and Gessaroli (1993) asserted that we (namely, sport and exercise psychology researchers) might not have the training and education required to understand the complexity of analyses that technology will make possible. Indeed, Schutz and Gessaroli echoed Tukey's (1986) sentiments that given the advances in technology that enable researchers to use complex statistical analyses, sport and exercise psychology researchers should increase their collaboration with statistical consultants. To this end, Zhu (2012) has called for sport and exercise psychometricians' in order to significantly improve the quality of measurement and validation practises in this emerging field. The results of this investigation indicate that researchers in sport and exercise psychology have yet to adopt modern validity theory frameworks. In line with Schutz and Gessaroli (1993) and Zhu (2012), we believe that further education on psychometric evaluation is necessary in order to bring validation research up to pace in sport and exercise psychology research. There are numerous publications that deal specifically with current validity theory or validation frameworks, many cited herein, from which researchers can design psychometric investigations with greater methodological rigour. One notable area that should be addressed is the issue of factor analysis. It appears as though factor analysis has become the *modus operandi* for examining score validity in sport and exercise psychology research. It is important that researchers (a) use factor analysis correctly (see Schutz and Gessaroli 1993 for an overview) and (b) realize that score validity extends beyond mere examination of factor structure (Zhu 2012).

Who Is Responsible for Accuracy?

There is no doubt that a researcher who conducts a validation investigation should be responsible for understanding and correctly applying validity theory. Of course, whether or not the investigation is publishable will be ascertained through the peer review process and editorial board stewardship. In particular, the onus falls on reviewers and journal editors to require evidence of score validity, require current and accurate terminology when referencing the validity of scores, and require validation papers that utilize and specify a validity theory or validation framework that substantiates the basis for their investigations. It is essential that researchers in sport and exercise psychology adopt more stringent protocols for validation and score validity such that any (and all) conclusions drawn from the proliferation of studies within this dynamic field offer greater insights for sport and health professionals (Hagger and Chatzisarantis 2009).

What Should Be Reported?

An investigation need not provide each source of evidence for score validity, but rather, report what is possible and meaningful given the context and sample under study (AERA et al. 1999). The validity evidence needed for scores of an instrument will depend on the claims that will be made with the scores (Kane 2013). For example, if one wishes to use scores from an instrument to make predictions about future behaviours, then predictive evidence is necessary (Kane 2013). If scores from the instrument are intended to be used for descriptive purposes only, predictive evidence may be unnecessary. To return to the referee self-efficacy instrument example, if scores from this instrument were intended to be used by league officials to select referees, this is a strong claim that requires a robust amount of validity evidence including (but not limited to) consequences and predictive evidence.

The *Standards* contend that "a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific use" (p. 17). If an investigator is altering an instrument, or employing an instrument in a sample/context that it was not originally intended for, it is the responsibility of that researcher to report evidence of score validity based on their sample before addressing the main aim of a given study (e.g., analysis of measurement invariance across samples as a precursor to examining between-group mean differences). Investigators should report some form of evidence of score validity for their instruments. This information, in combination with pre-existing evidence, would provide the basis for more substantive arguments pertinent to the validity of the inferences made from the instrument (AERA et al. 1999). Included in this validity argument may be the need to refine the instrument, the definition of the construct or other areas needed for further inquiry (AERA et al. 1999). Another important future direction for research is to conduct validity generalizability studies (Schmidt and Hunter 1977). Validity generalizability studies are similar to meta-analyses and should be conducted when there is sufficient data for a particular instrument.

Use a Validity Framework

Researchers who are conducting validation investigations should incorporate a validity framework into their investigations. By explicitly outlining a validity framework, researchers will be prevented from simply selecting the 'type' of validity they want to find and highlight what evidence is still needed (e.g., response processes). Researchers can use one of many validity frameworks outlined in the introduction; however, using the definitions and the validity framework outlined by the *Standards* could be one of the most viable and trustworthy routes to assess score validity. The *Standards* were created by committees that were made up of prominent scholars with expertise in measurement and validity theory. Furthermore, the *Standards* are advocated by the APA, which is an organization that many journals

in sport and exercise psychology adhere. If researchers and journal editors were to adopt conceptualizations of validity advocated within the *Standards*, research in sport and exercise psychology would be situated on a strong foundation, and as such, the credence of statements or conclusions derived from analysis of self-report instrumentation would likely be more defensible.

Limitations

It is important to acknowledge the results of this investigation cannot be generalizable to other investigations published in any journal other than *JSEP* between 2002 and 2012. Furthermore, the results of this investigation are limited to investigations that listed key words in the abstract and therefore cannot be generalized to investigations that did providence evidence of score validity but did not indicate it in their abstract.

Conclusion

Researchers conducting validation investigations within the discipline of sport and exercise psychology have generally not embraced modern validity theory or validation guidelines. Specifically, few researchers situated their research in a validation framework and fewer used language and terminology advocated by psychometricians. Results did reveal that many investigators provided multiple sources of validity evidence (e.g., content, internal structure, and relations to other variables) and used multiple samples to draw conclusions. Researchers who are interested in conducting validation investigations are encouraged to utilize a current validation framework that is based on modern validity theory (e.g., Messick's progressive matrix, the *Standards*). Basing a program of research on sound theoretical tenets will propel sport and exercise psychology research to the forefront of scientific inquiry.

References

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

Bartholomew, K. J., Ntoumanis, N., & Thøgersen-Ntoumani, C. (2010). The controlling interpersonal style in a coaching context: Development and initial validation of a psychometric scale. *Journal of Sport & Exercise Psychology, 32*, 193–216.

Boardley, I. D., & Kavussanu, M. (2007). Development and validation of the moral disengagement in sport scale. *Journal of Sport & Exercise Psychology, 29*, 608–628.

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

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 135–170). Charlotte: Information Age Publishing.

Brown, A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

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

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*, 397–410.

Dunn, J. G. H., Bouffard, M., & Rogers, W. T. (1999). Assessing item content-relevance in sport psychology scale-construction research: Issue and recommendations. *Measurement in Physical Education and Exercise Science, 3*, 15–36.

Freeman, P., Coffee, P., & Rees, T. (2011). The PASS-Q: The Perceived Available Support in Sport Questionnaire. *Journal of Sport & Exercise Psychology, 33*, 54–74.

Goodwin, L. D. (2002). Changing conceptions of measurement validity: An update on the new *Standards*. *Journal of Nursing Education, 41*, 100–106.

Gucciardi, D. (2011). The relationship between developmental experiences and mental toughness in adolescent cricketers. *Journal of Sport & Exercise Psychology, 33*, 370–393.

Hagger, M. S., & Chatzisarantis, N. L. D. (2009). Assumptions in research in sport and exercise psychology. *Psychology of Sport & Exercise, 10*, 511–519.

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.

Jones, M., Lane, A. M., Bray, S. R., Uphill, M., & Catlin, J. (2005). Development and validation of the Sport Emotion Questionnaire. *Journal of Sport & Exercise Psychology, 27*, 407–431.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Lonsdale, C., Hodge, K., & Rose, E. A. (2008). The Behavioral Regulation in Sport Questionnaire (BRSQ): Instrument development and initial validity evidence. *Journal of Sport & Exercise Psychology, 30*, 323–355.

Markland, D., & Tobin, V. (2004). A modification to the Behavioural Regulation in Exercise Questionnaire to include an assessment of amotivation. *Journal of Sport & Exercise Psychology, 26*, 191–196.

Messick, S. (1995). Validity of psychological assessment: Validations of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–774.

Morton, K. L., Barling, J., Rhodes, R. E., Mâsse, L. C., Zumbo, B. D., & Beauchamp, M. R. (2011). The application of transformational leadership theory to parenting: Questionnaire development implications for adolescent self-regulatory efficacy and life satisfaction. *Journal of Sport & Exercise Psychology, 33*, 688–709.

Myers, N. D., Feltz, D. L., Guillén, F., & Dithurbide, L. (2012). Development of, and initial validity evidence for, the Referee Self-Efficacy Scale: A multistudy report. *Journal of Sport & Exercise Psychology, 43*, 737–765.

Schmidt, F. L., & Hunter, J. E. (1977). Development and general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.

Schutz, R. W., & Gessaroli, M. E. (1993). Use, misuse, and disuse of psychometric in sport psychology research. In R. N. Singer, M. Murphy, & L. K. Tennant (Eds.), *Handbook of research on sport psychology* (pp. 901–917). New York: Macmillan.

Sebire, S. J., Standage, M., & Vansteenkiste, M. (2008). Development and validation of the goal content for exercise questionnaire. *Journal of Sport & Exercise Psychology, 30*, 353–377.

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

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte: Information Age Publishing.

Thompson, B. (1992). Two and one-half decades of leadership and measurement and devaluation. *Journal of Counseling and Development, 70*, 434–438.

Tukey, J. W. (1986). Sunset salvo. *The American Statistician, 40*, 72–76.

Williams, S. E., & Cumming, J. (2011). Measuring athlete imagery ability: The sport imagery ability questionnaire. *Journal of Sport & Exercise Psychology, 33*, 416–440.

Williams, S. E., Cumming, J., Ntoumanis, N., Nordin-Bates, S. M., Ramsey, R., & Hall, C. (2012). Further validation and development of the movement imagery questionnaire. *Journal of Sport & Exercise Psychology, 34*, 621–646.

Zhu, W. (1998). Comments on "development of a cadence curl-up test for college students" (Sparling, Millard-Stafford, & Snow, 1997): Concerns about validity and practicality. *Research Quarterly for Exercise and Sport, 69*, 308–310.

Zhu, W. (2012). Measurement practice in sport and exercise psychology: A historical comparative, and psychometric view. In G. Tenenbaum, R. C. Eklund, & A. Kamata (Eds.), *Measurement in sport and exercise psychology* (pp. 293–302). Champaign: Human Kinetics.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Pyschometrics, Vol. 26, pp. 45–79). Amsterdam: Elsevier.

Chapter 9
The Edinburgh Postnatal Depression Scale (EPDS): A Review of the Reported Validity Evidence

Hillary L. McBride, Rachel M. Wiens, Marvin J. McDonald, Daniel W. Cox, and Eric K.H. Chan

The perinatal period is characterized by significant changes in a woman's life. As her body grows and takes on a new shape, a woman can be filled with excitement, anticipation, fear, or any mixture of emotions. This occurs as she prepares for childbirth, copes with her changing identity, body, and relationships, and plans to meet her child (Lee 1995). Immediately following the birth of her child, the mother will receive a rush of hormones unlike any she will experience ever again, increasing her ability to breastfeed, form early attachment with her infant, and recover from the emotional and physical challenges of labour (Grattan 2011; Khajehei and Doherty 2012). In the time shortly following birth most women experience hormonally related mood sensitivity, including tearfulness, anxiety, and depressed mood, often referred to as the 'baby blues' (Bueno 2010). Bueno (2010) found these normal mood changes often peak between days three and five following birth, and gradually return to normal.

While this is the case for most women, between 10 and 15 % of most postpartum women experience a lasting depressed mood, with percentages of as high as 25 % in women with inadequate social support, low socioeconomic status, a history of or current mental illness, or women of adolescent age (Cox et al. 1987; Lanzi et al. 2009; Travis et al. 2012). Postpartum depression is defined as clinical depression

H.L. McBride (✉) • R.M. Wiens • M.J. McDonald
Trinity Western University, 7600 Glover Road, Langley, BC V2Y 1Y1, Canada
e-mail: Hillary.grams@mytwu.ca; rachel.wiens8@gmail.com; mcdonald@twu.ca

D.W. Cox
Counseling Psychology Program, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada

E.K.H. Chan, Ph.D.
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_9,

existing with the year following childbirth (Dennis et al. 2012). At least half of depressed mothers do not recover after the first postpartum year, putting at risk the wellbeing of their children (Cox et al. 1987). When exposed to maternal depression, newborns and infants demonstrate significant neurobehavioral dysregularities (Parcells 2010). Compared to infants of non-depressed mothers, those of mothers exhibiting depressive symptoms exhibit lower infant growth at 6 months, and may show behavioural disturbances at 3 years or cognitive defects at 4 years (Cogill et al. 1986; Travis et al. 2012; Wrate et al. 1985). These findings suggest the significant negative long-term impact of depression on postpartum mothers and their infants, and the necessity of appropriate and accurate screening to identify depression as early as possible during the postnatal period. The purposes of the present study were to review the validity evidence of the EPDS and examine the extent to which the reported evidence is in line with the modern view of validity as stated in the *Standards for Educational and Psychological Testing* (AERA et al. 1999).

Edinburgh Postnatal Depression Scale

The Edinburgh Postnatal Depression Scale (EPDS) was developed in 1987 by Cox, Holden, and Sagovsky, who identified existing depression screening scales were limited in their ability to assess potential depression in women during the postpartum period. For example, the Bedford and Foulds (1978) Anxiety and Depression Scale lacked validity when assessing pregnant women, as did the Beck Depression Inventory and the 30-item General Health Questionnaire from Goldberg et al. (1970). These scales were found to be inadequate when assessing postpartum women because the somatic symptoms present in an individual with depression occur naturally in pregnant women due to the physiological changes which occur following childbirth. In addition, community workers were reluctant to use such time-consuming questionnaires that appeared to lack face validity. For these reasons, a new scale was needed that would meet these challenges and adequately identify depression in postpartum women. In addition to being appropriate and simple to complete, the self-report scale needed to be acceptable to women who do not regard themselves as unwell. Such a scale must not require health workers to have any special knowledge or experience in the clinical diagnosis of depression, while also demonstrating satisfactory reliability and validity. In the development of the EPDS, three different scales were considered: the Irritability Depression and Anxiety Scale (IDA), the Hospital Anxiety and Depression Scale (HAD), and the Anxiety and Depression Scale of Bedford and Foulds (Bedford and Foulds 1978; Snaith et al. 1978; Zigmond and Snaith 1983). Twenty-one items were selected from these scales, including several items constructed by Cox et al. (1987), which were found to be appropriate for the detection of postnatal depression. The items were tested extensively with mothers of young babies, where 13 items were then selected as being more likely to detect postnatal depression. Seven items were

constructed by the researchers, while the other were adapted from the IDA and HAD. This led to the current and widely accepted version of the EPDS, created in 1987, which uses ten items to screen the test taker for possible depression (Cox and Holden 2003).

The EPDS is used internationally in routine postnatal care, and also used antenatally to assess a woman's likelihood of developing depressive symptoms following birth (Töreki et al. 2012). The test evaluates symptoms present in the 7 days prior to when the test is taken, and items are rated between zero and three, ranging from zero (symptom is not present) to three (symptom is severe) (Cox and Holden 2003). The test can be self-administered and any care provider can be trained to score and interpret the test, however this is most often done by the women's primary maternal care provider shortly following the birth of the baby during a routine postnatal visit (Segre et al. 2011). It typically takes approximately 5 min to administer and score, and will allow the care provider to make a referral for the woman to receive psychosocial support, or monitor the woman's mental health more closely (Cox and Holden 2003). Each item has a possible value of three points, where there is a zero value of the normal response, and a value of three for the extreme and highly symptomatic response. The total possible score on the test is 30. The acceptable cut off points for identifying potential depression range from 9 to 13, depending on the culture, language, and personal history of the participant (Logsdon et al. 2009; Pallant et al. 2006; Santos et al. 2007).

Reporting of Validity Evidence

A number of studies have revealed the inadequacy in the reporting of validity evidence in the academic literature. Barry et al. (2013) examined the frequency with which psychometric properties were reported in health education and behavioral journals. Of the 967 articles published in the seven journals reviewed between 2007 and 2010, an average of 67 % of the articles (between 40 and 93 %) did not report any validity evidence, while an average of 51 % (between 35 and 80 %) did not report any reliability evidence. In a review of the articles published in the *Journal of Counseling Psychology*, Meier and Davis (1990) found that only between 5 and 7 % of articles reviewed provided cited or sample reliability estimates respectively, while only 2 % of papers cited validity estimates or validity estimates for their samples.

Qualls and Moss (1996) found 47.5 % of instruments being used for assessment and testing purposes lacked reliability and validity information, where 41 % of instruments reported reliability information, and 31.7 % reported validity information. In 2009, Slaney, Tkatchouk, Gabriel and Maraun reported that of 368 articles published in 2004, 90.8 and 96.2 % addressed reliability and validity evidence, respectively. In 2010, Slaney and colleagues examined the frequency of reported reliability and validity information of measurement-oriented journals compared to a cross-section of research domains. In this article it was reported that of the

measurement-oriented journals, 87.6 % addressed reliability and 94.5 % addressed validity, while the cross-section of research domains addressed 95.8 and 94.4 % of reliability and validity information, respectively.

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 issue in the development and evaluation of the quality of a psychometric instrument. The AERA, APA, NCME (1999) *Standards for Educational and Psychological Testing* indicates five primary sources of validity evidence based on: (1) test content, (2) internal structure, (3) relationship to other variables, (4) response processes, and (5) consequences of testing.

Although Messick's (1989) modern view of validity has been published for over 25 years and has clearly influenced the AERA, APA, and NCME (1999) *Standards for Educational and Psychological Testing*, a number of studies examining the practices of validation in the psychology and education literature have shown that a number of sources of validity evidence are not presented. For instance, Cizek and colleagues found only 5 % and 1.7 % of the studies reporting discriminant validity and test consequences, respectively (Cizek et al. 2008, 2010). Similarly, a review of clinical assessment in internal medicine found that the reporting of response processes and consequences was absent (Auewarakul et al. 2005).

As a reminder, the objectives of the present study were to review the validity evidence of the EPDS and examine the extent to which the reported evidence is in line with the modern view of validity as stated in the *Standards for Educational and Psychological Testing* (AERA et al. 1999).

Method

Database Search

To identify articles useful for assessing the psychometric properties of the EPDS, articles from 1987 through 2013 were identified through the following databases: Academic Search Premier, America: History & Life, Biomedical Reference Collection, CINAHL, Communication & Mass Media, eBook Academic Collection (EBSCOhost), eBook Collection (EBSCOhost), E-Journals, ERIC, Funk & Wagnalls New World Encyclopedia, GreenFILE, Information Science & Technology, MAS Ultra, MEDLINE, Military & Government Collection, MLA Directory, Primary Search, PsychINFO, PsychBooks, Google Scholar and Sage Publication. During the primary search for the words "Edinburgh Postnatal Depression Scale", 4,142 sources were located. In addition to the words "Edinburgh Postnatal Depression Scale", we also searched for the words "valid", "validity", "validation" "psychometric", "reliability", "postpartum depression", "postnatal depression", and "depressive symptoms". As a result, 211 sources which appeared to be relevant

were selected for further review, which included a reading of the abstract by a member of the research team to determine if the source met criteria for inclusion. Of the 211 sources selected, 65 articles and 1 book were read in full by at least one member of the research team to further determine if the article would be appropriate for inclusion. The remaining articles that were not selected for inclusion in this study either did not address psychometric properties of the EPDS, or reviewed the validity and/or reliability evidence for non-postpartum related assessment. For example, some articles included validity evidence, but only when using the EPDS for assessing depression in women during menopause. From the thorough reading of those 65 sources, 57 articles were selected for analysis.

Inclusion Criteria and Coding

In order to be included in this study, the reviewed research had to address psychometric evidence for or against the EPDS. Of the articles reviewed, several were not written in English, or were validating the use of the EPDS on samples other than postpartum women (for example, men in the postpartum period, or women experiencing menopause). No studies were excluded based on publication date; all studies published since the creation of the scale in 1987 were included.

Each study selected was reviewed by two members of the research team and assessed for the following criteria: test content (face and content validity), associations (discriminant, convergent, concurrent and predictive validity), response processes, internal structure, and consequences. A coding sheet was created to chart the articles and book included, and which sources of validity and/or reliability evidence they contained. If a source addressed predictive validity, for example, an X was marked next to the study in the column listed as 'predictive validity', and the values given in the source for positive and negative predictive value were recorded. If a source did not include information about the predictive validity of the EPDS, that space in that column for the identified source was left blank.

Results

Of all studies surveyed, 57 were selected for further analysis. We first present an overview of the psychometric properties of the EPDS, followed by examining the extent to which the validity evidence of the EPDS as presented in the academic literature is in line with the modern view of validity as reflected in the most current version of the *Test Standards*.

Upon creation of the EPDS, validity was established through using a sample of 63 women, which showed that the 13 items clearly distinguished depressed and non-depressed women (Cox et al. 1987). A factor analysis showed that two items from the irritability subscale of the IDA, as well as one item concerning the

enjoyment of motherhood, created a separate "non-depression factor" (Cox et al. 1987, p. 783). Irritability was found to be a separate mood from depression and anxiety and for these reasons the three items were dropped. A validation study was then conducted on the 10-item scale.

The validation study on the 10-item EPDS was carried out on 84 mothers who had been evaluated at 6 weeks by visiting health care providers. These care providers were asked if the mothers were "depressed," "normal" or were having "problems" (Cox et al. 1987, p. 783). The scale was used to confirm the diagnosis of depression in women who were already suspected by their care worker as being depressed. Mothers were interviewed using Goldberg's Standardized Psychiatric Interview in their homes in order to determine if the scale would identify postnatal depression when administered in a domestic environment (Goldberg et al. 1970). The scale was then administered and results placed in a sealed envelope so that the interviewer remained blind to the score. In order to bypass any bias effect caused by the interviewer regarding participants as being "depressed," 12 non-depressed women were included in the sample. The Research Diagnostic Criteria (RDC) of Spitzer et al. (1975) was used as the criteria for diagnosis of a depressive illness. Validation of the 10-item EPDS was determined "by comparing EPDS scores with the RDC clinical diagnosis of depression" (Cox et al. 1987, p. 783).

The threshold score of 12/13 identified all of the 21 women with an RDC diagnosis of definite "Major Depressive Illness" and two of three women with probable "Major Depressive Illness" Cox et al. 1987, p. 784; Spitzer et al. 1975). Four of the 11 women with "Definite Minor Depression" were 'false positives' (Cox et al. 1987, p. 784). The sensitivity of the EPDS – proportion of RDC depressed women who were true positives – was 86 %. The specificity of the EPDS was found to be 78 % – non-depressed women who were true negatives. The positive predictive value of EPDS was 73 % – women who were above the EPDS threshold who met RDC criteria for depression. The failed detection of cases was reduced to fewer than 10 % with a cut off score of 9/10. Excluding the 12 women who had no previous identified problems increased the sensitivity to 85 %, the specificity to 77 %, and the positive predictive value to 83 %. The split-half reliability was found to be 0.88, with a standardized alpha-coefficient of 0.87.

Sensitivity to change in the severity of depression over time was analyzed by comparing EPDS scores at first interview and at an 11-week follow up interview. Mothers who were depressed, according to RDC criteria, at both interviews showed no significant difference, where mothers who were depressed at Interview 1, but not at Interview 2 showed a significant reduction on EPDS scores. The presence of a family member influenced EPDS scores; women were found to either exaggerate or minimize their symptoms. Three women who had a 'false positive score' and three of four who had a 'false negative' were not alone when interviewed. The EPDS was found to be useful in routine work of health workers in assisting to identify postnatal depression, although it does not substitute a clinical assessment.

Dimensions of the EPDS

The validity of the EPDS was questioned by Guedeney et al. (2000) when it was discovered that in a study of 87 postnatal women, three postpartum women who did in fact have major depression received false negatives on the test. During this assessment, a score of 10.5 was used as the cut off due to good sensitivity (0.80) and specificity (0.92). In each of the three women who received false negatives on the EPDS, somatic symptoms such as psychomotor retardation were present, but were not detected by the scale as it does not make use of a subscale addressing somatic symptoms. In these cases, dysphoric mood was displayed through flat affect, but never sadness as is measured by the scale used. The authors questioned the anxiety subscale used within the EPDS, and the lack of items addressing somatic symptoms of depression. This, however, was addressed by Cox et al. (1987) due to the confusion between the presence of normal somatic changes which occur postnatally and somatic symptoms occurring due to depression alone. Guedeney et al. (2000) suggested that this was due to author bias during the creation of the original EPDS, and that the scale is inadequate for detecting postpartum depression in women with a predominant profile of psychomotor retardation.

Reichenheim et al. (2011) agreed that the EPDS lacks a somatic subscale, however the researchers sought to determine whether the EPDS was one-dimensional, or multidimensional, in order to better understand if the EPDS test items lacked convergent validity and should be recreated with distinct tiers or subscales, or if it was appropriate for this scale to remain one-dimensional. Three factors emerged during the analysis as highly correlated, demonstrating that the model fit to be reasonably good: anhedonia, anxiety, and depression. The variance explained by the general factor accounted for 79.2 %, compared to 73.1 % of the variance accounted for by each three factors together. For this reason, the authors believe factors assessed with the EPDS should be considered a general factor. In light of the findings, the original 10 item EPDS appears to be well suited for screening postpartum depression in clinical practice.

Multidimensionality of the original EPDS was further supported in a study by Pallant et al. (2006) who determined, using Rasch analysis, a lack of fit to the model with significant item trait interaction. When two items were removed (items 7 and 8), non-significant item trait interaction occurred creating a fit to the model. The authors concluded that if the scale was reduced to eight items, it would be more robust psychometrically, but would need adjusted cut off rates; 7/8 and 9/10 were suggested. This, however, reduces scale items by 20 %, and decreases the opportunity to gain more information about symptoms from women taking the test, potentially decreasing validity.

The EPDS was also found to be a suitable screening tool for adolescent postpartum mothers (Logsdon et al. 2009). The subscales of anxiety and depression were detected. The anxiety subscale was a particularly to be a good fit with the feelings of abandonment, rejection and fear present in young mothers with

postpartum depression. It was, however, suggested that the adolescents' scores did not reach the adult threshold of 12 for detecting postpartum depression in adult women. This suggested further study is needed to determine which cut off may best detect postpartum depression in adolescent mothers, or symptomatology of adolescent postpartum depression.

Other Uses for the EPDS

The EPDS has been shown to be valid and reliable when screening for postpartum depression, however it has also been effective when screening for depression during pregnancy (Bergink et al. 2011). Further, the EPDS was shown to be helpful for assessing a woman's likelihood of developing postpartum depression, and giving the maternity care provider information about depressive symptoms already occurring. For this reason Bergink et al. (2011) have suggested the EPDS be renamed the Edinburgh Depression Scale, to be inclusive of the entire perinatal period. In this study, test-retest correlation coefficients were high (Spearman correlation coefficient was 0.55–0.63), and concurrent validity was also found to be high when compared to the Symptom Checklist – 90 items (SCL-90) anxiety and somatization subscales. The scale also had high internal consistency as assessed by Cronbach's alpha ($\alpha = 0.82–0.84$). Cut off values of 10 during the second and third trimester, and 11 during the first trimester have the best combination of sensitivity, specificity and positive predictive values.

Validity Evidence of the EPDS and the *Test Standards*

In the 57 studies included in the present analysis, five (8.8 %) addressed face validity of the EPDS, and six (10.5 %) discussed content validity of the scale. When considering the associations of validity, three studies (5.3 %) discussed discriminant validity, 15 (26.3 %) addressed convergent validity, 27 (47.4 %) reported on concurrent validity, and 29 (51 %) discussed predictive validity. Only one study (1.8 %) mentioned response processes, and two studies (3.5 %) mentioned test consequences. Sixteen (28.1 %) studies addressed the internal structure of the EPDS and 30 (52.6 %) mentioned internal consistency.

Content Validity

Only six (10.5 %) of the research studies that were examined addressed the content validity of the EPDS (Benvenuti et al. 1999; Guedeney et al. 2000; Hanlon et al. 2008; Tesfaye et al. 2009; Vivilaki et al. 2009; Wang et al. 2009). Content validity

was reported to be adequate in studies where the EPDS was translated into other languages (Benvenuti et al. 1999; Leonardou et al. 2009). The adequate level of content validity of the EPDS suggests that the items on the scale depict a thorough representation of the construct of postnatal depression.

In a study by Benvenuti and colleagues (1999), the EPDS was translated from English to Italian, where authors assessed each item for cultural equivalence. The items on the Italian version of the EPDS were found to achieve the same content validity as the original English version; however authors did not indicate the method through which they verified content validity.

Another study examining content validity is the article by Guedeney et al. (2000) which examines three cases of 'false negatives'. Through an examination of the literature, the authors address how existing tests assessing or screening for depression "reflect different emphasis in item content of the questionnaires which, in turn, reflects different notions of 'depression' held by the designers of the instruments" (p. 110). Guedeney and colleagues identify that the EPDS is designed from the theory of depression focused on the anhedonic symptoms, as opposed to the psychomotor retardation. It was for this reason that the EPDS was not able to identify the participants as potential cases of depression.

In the study by Hanlon et al. (2008), two independent sets of Ethiopian physicians agreed upon the test items, which were then verified by two experienced Ethiopian psychiatrists who also had extensive experience with community screening tools. In a comparable study by Tesfaye et al. (2009) for the use of the EPDS in Ethiopia, a panel of translators, psychiatrists and psychologists assembled to ensure the content validity of the EPDS, and make changes accordingly. Similarly, the study of the Greek translation of the EPDS by Vivilaki et al. (2009) used a panel of midwives to assess the content validity of the EPDS. Unlike other studies, this panel identified both anxiety and depression as components of the EPDS, which they believed matched the symptoms of postpartum depression.

In the final study exploring content validity (Wang et al. 2009) the EPDS was evaluated by a panel of psychiatrists, psychiatric nurses, obstetricians and obstetrical nurses with extensive knowledge of the concepts and instruments. The panel evaluated each item on a Likert scale assessing the applicability of the expression and content of the item, and the results were then used to calculate the Content Validity Index (CVI). The final CVI was 0.93, which demonstrated a satisfactory consensus among the evaluation panel.

Association with Other Variables

Only three articles (5.3 %) addressed discriminant validity (Lau et al. 2010; Reichenheim et al. 2011; Wang et al. 2009). Fifteen articles (26.3 %) addressed convergent validity; thus the majority of these studies correlated the EPDS with other tests that measured postnatal depression in order to establish validity (Berle et al. 2003; Brouwers et al. 2001; De Bruin et al. 2004; Guedeney and Fermanian

1998; Hanlon et al. 2008; Lau et al. 2010; Logsdon et al. 2009; Mazhari and Nakhaee 2007; Montazeri et al. 2007; Phillips et al. 2009; Reichenheim et al. 2011; Small et al. 2007; Vivilaki et al 2009; Wang et al. 2009; Yawn et al. 2009). The EPDS was correlated with such measures as the General Health Questionnaire (Shelton and Herrick 2009) and the Self-Report Questionnaire (Santos et al. 2007). The EPDS was determined to be a valid measure of postnatal depression. Many research studies also addressed concurrent validity (47.4 %) and predictive validity (51 %). The concurrent validity of the EPDS was questioned by researchers due to false negatives and the test's inability to detect somatic symptoms (Guedeney et al. 2000; Reichenheim et al. 2011). The predictive validity of the EPDS was established as it was able to distinguish depressed from non-depressed mothers (Dennis 2004) as well as predict and screen for depression in participants (Benvenuti et al. 1999; Garcia-Esteve et al. 2003; Santos et al. 2007).

Response Processes

Only one (1.8 %) study addressed response processes associated with the EPDS (Godderis et al. 2009). The researchers sought to understand both the cognitive processes influencing participants' responses to the EPDS, as well as how the participants understood the individual questions on the EPDS. The cognitive aspects of survey methodology (CASM) was used to evaluate the responses processes of the participants, and the cognitive interviewing focused on how the participants understood, interpreted, and answered the questions on the EPDS. The cognitive interviews incorporated probing questions about each item on the EPDS. The authors provided an example of this technique using item one: "I have been able to laugh and see the funny side of things". Through cognitive interviewing, the researchers used this question to comprehend how the phrase "the funny side of things" may be interpreted by the participants. Part of the interviewing involved the participant reading the questions and her answers aloud. Following this, the interviewer would ask the participant if she had any additional comments, responses, or questions. The interviewer would then ask the participant what she thought was meant by the particular phrase that is being tested ("the funny side of things"), and ask the participant to provide examples. The researchers asked probing questions for each item on the EPDS, and asked all the participants about their reactions to completing the scale.

Internal Structure

Sixteen (28.1 %) studies examined the internal structure of the EPDS. Although there is evidence supporting the multidimensional measurement structure of the EPDS as described in the Dimensions of the EPDS section of this chapter (Chabrol

and Teissedre 2004; Kheriabadi et al. 2012; Pallant et al. 2006; Pop et al. 1992; Reichenheim et al. 2011), some have argued that the instrument's structure should be unidimensional. The unidimensionality of the EPDS has also been explored. De Bruin et al. (2004) examined the factor structure of the EPDS through confirmatory factor analysis. Although items 1,2,6,7,8,9, and 10 are often seen as indicators of depressive feelings, while items 3, 4, and 5 are seen as indicators of anxiety, the researchers determined that "a single common factor underlies responses to the ten items" (p. 119) and displayed that "the two-factor model is not superior to the one-factor model" (p. 119). This is similar to the results of a study by Toreki et al. (2012) who identified three distinct factors. The researchers determined that these factors were not clinically significant, and neither did they reflect a multidimensional structure of the EPDS. It appears that there is still conflicting evidence on the number of dimensions the EPDS possesses.

Consequences

The consequences of the EPDS were only addressed in two (3.5 %) of the research articles that were examined (Downie et al. 2003; Krantz et al. 2008). This small number is significant considering the arguments made by the authors of both studies, which highlighted ethical issues and follow up care provided to women with high scores.

Krantz and colleagues (2008) suggest that the EPDS does not function well as a routine screening instrument and therefore increases the likelihood of false negative or positives. Authors determined that this poses a threat to the ethical principal of beneficence and autonomy, and can potentially cause more harm than good when categorizing women incorrectly. It also encourages a dualistic and reductionist understanding of mental health, which fails to consider the variety of experiences of health and distress.

Downie and colleagues (2003) had different conclusions, suggesting that the EPDS was in fact a useful screening tool. However, their study indicated that the care provider administering the test also influenced test consequences. Women with high scores had contact with a nurse who had administered the test and was able to provide follow up care. In some cases, these were women who were unlikely to seek out additional support, but had it provided to them due to their contact with the nurse. In other cases, the nurses' role was more hindering, as they failed to provide referrals to women who had received high scores. In order for the EPDS to be effective as a screening tool, care provider follow up and access to services are essential.

Discussion

It was determined through the articles reviewed in this study that there is sufficient evidence to support the validity and reliability of the EPDS as a screening tool for postnatal depression. In some cases, the EPDS was also supported for other purposes such as screening for antepartum depression, postpartum depression in men, and the prediction of postpartum depression. The validity, positive and negative predictive values, sensitivity, and specificity of the EPDS appears to differ slightly depending on the sample population, the cut off score selected, the point during the perinatal period at which the participant had taken the test, the language of the test, and the participant's age. Some evidence was shown that participants taking the test in the presence of the test administrator had different scores than when they took the test with a family member present. Although the test has been found suitable for screening for postpartum depression and other types of perinatal depression, other symptoms not assessed for by the EPDS are useful for assessing the severity of a woman's depression. This may be the case if a woman is experiencing severe somatic symptoms, such as psychomotor retardation: her test score may be low, but it may have taken 45 min to complete the scale, which usually takes about 5 min to complete.

Each article included various aspects of the reliability and validity evidence of the EPDS. Of the 57 included studies, many addressed concurrent validity of the scale (47.4 %), predictive validity (51 %), and internal consistency (52.6 %). Fewer studies addressed convergent validity (26.3 %) and internal structures (28.1 %). There was a paucity of information discussing face validity (8.8 %), content validity (10.5 %), discriminant validity (5.3 %), response processes (1.8 %) and test consequences (3.5 %).

Of the 57 studies included in this study for addressing the psychometric properties of the EPDS, only two of these studies discussed consequences. This is consistent with the results of Cizek et al. (2010), who found that only 0.7 % of tests examined addressed consequential validity. From the results of their study, the researchers concluded that consequential validity reporting is not up to standards of modern validity theory. Cizek and colleagues further noted that it would be expected that consequential validity be reported more than it appears to be in test evaluation, particularly because the notion of consequential validity as evidence in validity theory has been known for at least two decades.

Only three studies included addressed discriminant validity, and only five and six studies mentioned face validity and content validity, respectively. According to the *Test Standards*, validity must be established by providing multiple examples of validity evidence rather than using a single method; the lack of studies addressing the discriminant, face, and content validity may not be adequate to determine the validity of the EPDS. The studies included also validated varying versions of the EPDS, on various samples, which may have affected the results in supporting the validity and reliability of the study. These limitations may be due to the limited

number of articles retrieved from the data bases during the search, or a lack of literature for the validation of the EPDS.

Additionally, some studies did not report statistical figures when discussing the validity evidence for the EPDS. According to the *Test Standards* (AERA et al. 1999), emphasis on quality of validity evidence is more important than quantity; it is thus unfortunate that researchers in many of the studies did not provide statistical evidence when discussing the various types of validity evidence. Reporting specific statistical results would provide greater clarity and support for validity.

The *Test Standards* (1999) define validity as "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (p. 9). Though the majority of the articles included in the study had adequate evidence to support the validity of the EPDS, there were many articles that demonstrated lack of evidence regarding test consequences, face validity, content validity, and discriminant validity. Response processes and consequences are the additional sources of validity evidence needed to strengthen the score inferences (Hubley and Zumbo 2011, 2013; Zumbo 2007, 2009) made from EPDS.

References[1]

*Alvarado-Esquivel, C., Sifuentes-Alvarez, A., Salas-Martinez, C., & Martinez-Garcia, S. (2006). Validation of the Edinburgh Postpartum Depression Scale in a population of puerperal women in Mexico. *Clinical Practice and Epidemiology in Mental Health, 2*, 33–38. doi:10.1186/1745-0179-2-33.

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

Auewarakul, C., Downing, S. M., Jaturatamrong, U., & Praditsuwan, R. (2005). Sources of validity evidence for an internal medicine student evaluation system: An evaluative study of assessment methods. *Medical Education, 39*(3), 276–283. doi:10.1111/j.1365-2929.2005.02090.x.

*Aydin, N., Inandi, T., Yigit, A., & Hodoglugil, N. N. S. (2004). Validation of the Turkish version of the Edinburgh Postnatal Depression Scale among women within their first postpartum year. *Social Psychiatry, 39*, 483–486. doi:10.1007/s00127-004-0770-4.

*Banerjee, N., Banerjee, A., Kriplani, A., Saxena, S., & Banerjee, A. (2000). Evaluation of Postpartum depression using the Edinburgh Postnatal Depression Scale in a rural community in India. *International Journal of Social Psychiatry, 46*(1), 74–75. doi:10.1177/002076400004600109.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2013). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*. doi:10.1177/1090198113483139.

Bedford, A., & Foulds, G. A. (1978). *Delusions-symptoms-states inventory state of anxiety and depression (manual)* (pp. 1–14). Berkshire: NFER.

[1] References marked with an asterisk indicate studies included in this review.

*Benjamin, D., Chandramohan, A., Annie, I. K., Prasad, J., & Jacob, K. S. (2005). Validation of the Tamil version of the Edinburgh Post-partum Depression Scale. *The Journal of Obstetrics and Gynecology of India, 55*(3), 241–243.

*Benvenuti, P., Ferrara, M., Niccolai, C., Valoriani, V., & Cox, J. (1999). The Edinburgh Postnatal Depression Scale: Validation for an Italian sample. *Journal of Affective Disorders, 53*, 137–141.

Bergink, V., Kooistra, L., Lambregtse-van den Berg, M., Wignen, H., Bunevicius, R., van Baar, A., & Pop, V. (2011). Validation of the Edinburgh Depression Scale during pregnancy. *Journal of Psychosomatic Research, 70*, 385–389. doi:10.1016/j.jpsychores.2010.07.008.

*Berle, J. O., Aarre, T. F., Mykletun, A., Dahl, A. A., & Holsten, F. (2003). Screening for postnatal depression: Validation of the Norwegian version of the Edinburgh Postnatal Depression Scale, and assessment of risk factors for postnatal depression. *Journal of Affective Disorders, 76*, 151–156. doi:10.1016/S0165-0327(02)00082-4.

*Boyce, P., Stubbs, J., & Todd, A. (1993). The Edinburgh Postnatal Depression Scale: Validation for an Australian sample. *Australian and New Zealand Journal of Psychiatry, 27*(3), 472–476.

*Brouwers, E. P., van Baar, A. L., & Pop, V. J. (2001). Does the Edinburgh Postnatal Depression Scale measure anxiety? *Journal of Psychosomatic Research, 51*(5), 659–663.

Bueno, J. (2010). Life after birth. *Therapy Today, 21*(4), 18–21.

*Bunevičius, A., Kusminskas, L., & Bunevičius, R. (2009). Validation of the Lithuanian version of the Edinburgh Postnatal Depression Scale. *Medicina (Kaunas, Lithuania), 45*(7), 544–548.

*Carpiniello, B., Pariante, C. M., Serri, F., Costa, G., & Carta, M. G. (1997). Validation of the Edinburgh Postnatal Depression Scale in Italy. *Journal of Psychosomatic Obstetrics & Gynecology, 18*(4), 280–285.

*Chabrol, H., & Teissedre, F. (2004). Relation between Edinburgh Postnatal Depression Scale scores at 2–3 days and 4–6 weeks postpartum. *Journal of Reproductive and Infant Psychology, 22*(1), 33–39. doi:10.1080/02646830310001643067.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397–412. doi:10.1177/0013164407310130.

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*(5), 732–774. doi:10.1177/0013164410379323.

Cogill, S. R., Caplan, H. L., Alexandra, H., Robson, K. M., & Kumar, R. (1986). Impact of maternal postnatal depression on cognitive development of young children. *British Medical Journal, 292*, 1165–1167.

Cox, J., & Holden, J. (2003). *Perinatal mental health: A guide to the Edinburgh Postnatal Depression Scale (EPDS)*. London: Gaskell.

*Cox, J. L., Holden, J. M., & Sagovsky, R. (1987). Detection of postnatal depression: Development of the 10-item Edinburgh Postnatal Depression Scale. *The British Journal of Psychiatry, 150*(6), 782–786.

*De Bruin, G. P., Swartz, L., Tomlinson, M., Cooper, P. J., & Molteno, C. (2004). The factor structure of the Edinburgh Postnatal Depression Scale in a South African peri-urban settlement. *South African Journal of Psychology, 34*(1), 113–121. doi:10.1177/008124630403400107.

*Dennis, C. L., Janssen, P. A., & Singer, J. (2004). Identifying women at-risk for postpartum depression in the immediate postpartum period. *Acta Psychiatrica Scandinavica, 110*(5), 338–346. doi:10.1111/j.1600-0447.2004.00337.x.

Dennis, C., Heaman, M., & Vigod, S. (2012). Epidemiology of postpartum depressive symptoms among Canadian women: Regional and national results from a cross-sectional survey. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie, 57*(9), 537–546.

*Downie, J., Wynaden, D., McGowan, S., Juliff, D., Axten, C., Fitzpatrick, L., Ogilvie, S., & Painter, S. (2003). Using the Edinburgh Postnatal Depression Scale to achieve best practice standards. *Nursing and Health Sciences, 5*, 283–287.

*Felice, E., Saliba, J., Grech, V., & Cox, J. (2006). Validation of the Maltese version of the Edinburgh postnatal depression scale. *Archives of Women's Mental Health, 9*(2), 75–80. doi:10.1007/s00737-005-0099-3.

*Garcia-Esteve, L., Ascaso, C., Ojuel, J., & Navarro, P. (2003). Validation of the Edinburgh postnatal depression scale (EPDS) in Spanish mothers. *Journal of Affective Disorders, 75*(1), 71–76. doi:10.1016/S0165-0327(02)00020-4.

*Gausia, K., Fisher, C., Algin, S., & Oosthuizen, J. (2007). Validation of the Bangla version of the Edinburgh Postnatal Depression Scale for a Bangladeshi sample. *Journal of Reproductive and Infant Psychology, 25*(4), 308–315. doi:10.1080/02646830701644896.

*Ghubash, R., Abou-Saleh, M. T., & Daradkeh, T. K. (1997). The validity of the Arabic Edinburgh Postnatal Depression Scale. *Social Psychiatry and Psychiatric Epidemiology, 32*(8), 474–476.

*Godderis, R., Adair, C. E., & Brager, N. (2009). Applying new techniques to an old ally: A qualitative validation study of the Edinburgh Postnatal Depression Scale. *Women and Birth, 22*(1), 17–23. doi:10.1016/j.wombi.2008.10.002.

Goldberg, D. P., Cooper, B., Eastwood, M. R., Kedward, H. B., & Shepherd, M. (1970). A standardized psychiatric interview for use in community surveys. *British Journal of Preventive & Social Medicine, 24*(1), 18–23.

Grattan, D. (2011). A mother's brain knows. *Journal of Neuroendocrinology, 23*(11), 1188–1189. doi:10.1111/j.1365-2826.2011.02175.x.

*Guedeney, N., & Fermanian, J. (1998). Validation study of the French version of the Edinburgh Postnatal Depression Scale (EPDS): New results about use and psychometric properties. *European Psychiatry, 13*(2), 83–89.

*Guedeney, N., Fermanian, J., Guelfi, J. D., & Kumar, R. C. (2000). The Edinburgh Postnatal Depression Scale (EPDS) and the detection of major depressive disorders in early postpartum: Some concerns about false negatives. *Journal of Affective Disorders, 61*, 107–122.

*Hanlon, C., Medhin, G., Alem, A., Araya, M., Abdulahi, A., Hughes, M., Tesfaye, M., Wondimagegn, D., Patel, V., & Prince, M. (2008). Detecting perinatal common mental disorders in Ethiopia: Validation of the self-reporting questionnaire and Edinburgh Postnatal Depression Scale. *Journal of Affective Disorders, 108*(3), 251–262. doi:10.1016/j.jad.2007.10.023.

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.

*Jadresic, E., Araya, R., & Jara, C. (1995). Validation of the Edinburgh postnatal depression scale (EPDS) in Chilean postpartum women. *American Journal of Obstetrics and Gynecology, 16*, 187–191.

Khajehei, M., & Doherty, M. (2012). Childbirth in pleasure and ecstasy: A fountain of hormones and chemicals. *International Journal of Childbirth Education, 27*(3), 73–80.

*Kheirabadi, G. R., Maracy, M. R., Akbaripour, S., & Masaeli, N. (2012). Psychometric properties and diagnostic accuracy of the Edinburgh Postnatal Depression Scale in a sample of Iranian women. *Iranian Journal of Medical Sciences, 37*(1), 32.

*Krantz, I., Eriksson, B., Lundquist-Persson, C., Ahlberg, B. M., & Nilstun, T. (2008). Screening for postpartum depression with the Edinburgh Postnatal Depression Scale (EPDS): An ethical analysis. *Scandinavian Journal of Public Health, 36*(2), 211–216. doi:10.1177/1403494807085392.

Lanzi, R. G., Bert, S. C., & Jacobs, B. K. (2009). Depression among a sample of first-time adolescent and adult mothers. *Journal of Child and Adolescent Psychiatric Nursing, 22*(4), 194–202. doi:10.1111/j.1744-6171.2009.00199.x.

*Lau, Y., Wang, Y., Yin, L., Chan, K. S., & Guo, X. (2010). Validation of the mainland Chinese version of the Edinburgh Postnatal Depression Scale in Chengdu mothers. *International Journal of Nursing Studies, 47*(9), 1139–1151. doi:10.1016/j.ijnurstu.2010.02.005.

*Lawrie, T. A., Hofmeyr, G. J., de Jager, M., & Berk, M. (1998). Validation of the Edinburgh Postnatal Depression Scale on a cohort of South African women. *South African Medical Journal, 88*(10), 1340–1344.

Lee, R. E. (1995). Women look at their experience of pregnancy. *Infant Mental Health Journal, 16* (3), 192–205.

*Lee, D. T., Yip, S. K., Chiu, H. F., Leung, T. Y., Chan, K. P., Chau, I. O., Leung, H. C., & Chung, T. K. (1998). Detecting postnatal depression in Chinese women: Validation of the Chinese version of the Edinburgh Postnatal Depression Scale. *The British Journal of Psychiatry, 172* (5), 433–437.

*Leonardou, A. A., Zervas, Y. M., Papageorgiou, C. C., Marks, M. N., Tsartsara, E. C., Antsaklis, A. A., & Soldatos, C. R. (2009). Validation of the Edinburgh Postnatal Depression Scale and prevalence of postnatal depression at two months postpartum in a sample of Greek mothers. *Journal of Reproductive and Infant Psychology, 27*(1), 28–39. doi:10.1080/02646830802004909.

*Leverton, T. J., & Elliott, S. A. (2000). Is the EPDS a magic wand?: A comparison of the Edinburgh Postnatal Depression Scale and health visitor report as predictors of diagnosis on the present state examination. *Journal of Reproductive and Infant Psychology, 18*(4), 279–296. doi:10.1080/713683048.

*Logsdon, M. C., Usui, W. M., & Nering, M. (2009). Validation of Edinburgh Postnatal Depression Scale for adolescent mothers. *Archives of Women's Mental Health, 12*, 433–440.

*Mazhari, S., & Nakhaee, N. (2007). Validation of the Edinburgh postnatal depression scale in an Iranian sample. *Archives of Women's Mental Health, 10*(6), 293–297. doi:10.1007/s00737-007-0204-x.

*McCoy, S. J. B., Beal, J. M., Payton, M. E., Stewart, A. L., DeMers, A. M., & Watson, G. H. (2005). Correlations of visual analog scales with Edinburgh Postnatal Depression Scale. *Journal of Affective Disorders, 86*(2), 295–297. doi:10.1016/j.jad.2005.01.001.

Meier, S. T., & David, S. R. (1990). Trends in reporting psychometric properties of scales used in counselling psychology research. *Journal of Counseling Psychology, 37*(1), 113–115. doi:10.1037/0022-0167.37.1.113.

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

*Milgrom, J., Ericksen, J., Negri, L., & Gemmill, A. W. (2005). Screening for postnatal depression in routine primary care: Properties of the Edinburgh Postnatal Depression Scale in an Australian sample. *Australian and New Zealand Journal of Psychiatry, 39*(9), 833–839. doi:10.1080/j.1440-1614.2005.01660.x.

*Milgrom, J., Mendelsohn, J., & Gemmill, A. W. (2011). Does postnatal depression screening work? Throwing out the bathwater, keeping the baby. *Journal of Affective Disorders, 132*(3), 301–310. doi:10.1016/j.jad.2010.09.031.

*Montazeri, A., Torkan, B., & Omidvari, S. (2007). The Edinburgh Postnatal Depression Scale (EPDS): Translation and validation study of the Iranian version. *BMC Psychiatry, 7*(1), 11. doi:10.1186/1471-244X-7-11.

*Murray, L., & Carothers, A. D. (1990). The validation of the Edinburgh Post-natal Depression Scale on a community sample. *The British Journal of Psychiatry, 157*(2), 288–290.

*Navarro, P., Ascaso, C., Garcia-Esteve, L., Aguado, J., Torres, A., & Martín-Santos, R. (2007). Postnatal psychiatric morbidity: A validation study of the GHQ-12 and the EPDS as screening tools. *General Hospital Psychiatry, 29*(1), 1–7. doi:10.1016/j.genhosppsych.2006.10.004.

*Pallant, J., Miller, R., & Tennant, A. (2006). Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. *BMC Psychiatry, 6*(28), 1–10.

Parcells, D. A. (2010). Women's mental health nursing: Depression, anxiety and stress during pregnancy. *Journal of Psychiatric and Mental Health Nursing, 17*(9), 813–820. doi:10.1111/j.1365-2850.2010.01588.x.

*Phillips, J., Charles, M., Sharpe, L., & Matthey, S. (2009). Validation of the subscales of the Edinburgh Postnatal Depression Scale in a sample of women with unsettled infants. *Journal of Affective Disorders, 118*(1), 101–112. doi:10.1016/j.jad.2009.02.004.

*Pitanupong, J., Liabsuetrakul, T., & Vittayanont, A. (2007). Validation of the Thai Edinburgh Postnatal Depression Scale for screening postpartum depression. *Psychiatry Research, 149*(1), 253–259. doi:10.1016/j.psychres.2005.12.011.

*Pop, V. J., Komproe, I. H., & Van Son, M. J. (1992). Characteristics of the Edinburgh post natal depression scale in The Netherlands. *Journal of Affective Disorders, 26*(2), 105–110.

Qualls, A. L., & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. *Educational and Psychological Measurement, 56*(2), 209–214. doi:10.1177/0013164496056002002.

*Regmi, S., Sligl, W., Carter, D., Grut, W., & Seear, M. (2002). A controlled study of postpartum depression among Nepalese women: Validation of the Edinburgh Postpartum Depression Scale in Kathmandu. *Tropical Medicine & International Health, 7*(4), 378–382.

*Reichenheim, M., Morales, C. L., Oliveira, A., & Lobato, G. (2011). Revisiting the dimensional structure of the Edinburg Postnatal Depression Scale (EPDS): Empirical evidence for a general factor. *BMC Medical Research Methodology, 11*(93), 1–12.

*Santos, I., Matijasevich, A., Tavares, B., da Cruz Lima, A., Riegel, R., & Lopes, B. (2007). Comparing validity of Edinburg scale and SRQ20 in screening for post-partum depression. *Clinical Practice and Epidemiology in Mental Health, 3*(18), 1–5.

Segre, L., Brock, R., O'Hara, M., Gorman, L., & Engeldinger, J. (2011). Disseminating perinatal depression screening as a public health initiative: A Train-the-Trainer approach. *Maternal and Child Health Journal, 15*(6), 814–821. doi:10.1007/s10995-010-0644-1.

*Shelton, J., & Herrick, K. (2009). Comparison of scoring methods and thresholds of the General Health Questionnaire-12 with the Edinburgh Postnatal Depression Scale in English women. *Public Health, 123*(12), 789–793. doi:10.1016/j.puhe.2009.09.012.

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., & Maraun, M. D. (2009). Psychometric assessment and reporting practice: Incongruence between theory and practice. *Journal of Psychoeducational Assessment, 27*(6), 465–476. doi:10.1177/0734282909335781.

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., Ferguson, L. P., Knudsen, J. R., & Legere, J. C. (2010). A review of psychometric assessment and reporting practices: An examination of measurement-oriented versus non-measurement-oriented domains. *Canadian Journal of School Psychology, 25*(3), 246–259. doi:10.1177/0829573510375549.

*Small, R., Lumley, J., Yelland, J., & Brown, S. (2007). The performance of the Edinburgh Postnatal Depression Scale in English speaking and non-English speaking populations in Australia. *Social Psychiatry and Psychiatric Epidemiology, 42*(1), 70–78. doi:10.1007/s00127-006-0134-3.

Snaith, R. P., Constantopoulos, A. A., Jardine, M. Y., & McGuffin, P. (1978). A clinical scale for the self-assessment of irritability. *The British Journal of Psychiatry, 132*(2), 164–171.

Spitzer, R. L., Endicott, J., Robins, E., Kuriansky, J., & Gurland, B. (1975). Preliminary report of the reliability of research diagnostic criteria (RDC) applied to psychiatric case records. In A. Sudilovsky & R. Beer (Eds.), *Prediction in psychopharmacology: Preclinical and clinical correlates*. San Diego: Raven.

*Teissèdre, F., & Chabrol, H. (2004). Detecting women at risk for postnatal depression using the Edinburgh Postnatal Depression Scale at 2 to 3 days postpartum. *Canadian Journal of Psychiatry, 49*(1), 51–54.

*Teng, H. W., Hsu, C. S., Shih, S. M., Lu, M. L., Pan, J. J., & Shen, W. W. (2005). Screening postpartum depression with the Taiwanese version of the Edinburgh Postnatal Depression Scale. *Comprehensive Psychiatry, 46*(4), 261–265. doi:10.1016/j.comppsych.2004.10.003.

*Tesfaye, M., Hanlon, C., Wondimagegn, D., & Alem, A. (2009). Detecting postnatal common mental disorders in Addis Ababa, Ethiopia: Validation of the Edinburgh postnatal depression scale and Kessler scales. *Journal of Affective Disorders, 122*(1), 102–108. doi:10.1016/j.jad.2009.06.020.

Töreki, A., Andó, B., Keresztúri, A., Sikovanyecz, J., Dudas, R., Janka, Z., Kozinszky, Z., & Pál, A. (2012). The Edinburgh Postnatal Depression Scale: Translation and antepartum validation for a Hungarian sample. *Midwifery, 29*(4), 308–315. doi:10.1016/j.midw.2012.01011.

Traviss, G. D., West, R. M., & House, A. O. (2012). Maternal mental health and its association with infant growth at 6 months in ethnic groups: Results from the Born-in-Bradford Birth Cohort Study. *PLoS One, 7*(2), e30707. doi:10.1371/journal.pone.0030707.

*Vivilaki, V. G., Dafermos, V., Kogevinas, M., Bitsios, P., & Lionis, C. (2009). The Edinburgh Postnatal Depression Scale: Translation and validation for a Greek sample. *BMC Public Health, 9*(1), 329. doi:10.1186/1471-2458-9-329.

*Wang, Y., Guo, X., Lau, Y., Chan, K. S., Yin, L., & Chen, J. (2009). Psychometric evaluation of the mainland Chinese version of the Edinburgh Postnatal Depression Scale. *International Journal of Nursing Studies, 46*(6), 813–823. doi:10.1016/j.ijnurstu.2009.01.010.

*Werrett, J., & Clifford, C. (2006). Validation of the Punjabi version of the Edinburgh Postnatal Depression Scale (EPDS). *International Journal of Nursing Studies, 43*(2), 227–236. doi:10.1016/j.ijnurstu.2004.12.007.

*Wickberg, B., & Hwang, C. P. (1996). The Edinburgh Postnatal Depression Scale: Validation on a Swedish community sample. *Acta Psychiatrica Scandinavica, 94*(3), 181–184.

*Wojcicki, J. M., & Geissler, J. (2013). The Spanish translation of the Edinburgh Postnatal Depression Scale and the use of the word “desgraciada”. *Transcultural Psychiatry, 50*(1), 152–154. doi:10.1177/1363461512475276.

Wrate, R. M., Rooney, A. C., Thomas, P. F., & Cox, J. L. (1985). Postnatal depression and child development: A three-year follow-up study. *The British Journal of Psychiatry, 146*, 622–627. doi:10.1192/bjp.146.6.622.

*Yawn, B. P., Pace, W., Wollan, P. C., Bertram, S., Kurland, M., Graham, D., & Dietrich, A. (2009). Concordance of Edinburgh Postnatal Depression Scale (EPDS) and Patient Health Questionnaire (PHQ-9) to assess increased risk of depression among postpartum women. *The Journal of the American Board of Family Medicine, 22*(5), 483–491. doi:10.3122/jabfm.2009.05.080155.

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica, 67*(6), 361–370.

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: Information Age Publishing.

Chapter 10
Validity Theory and Validity Evidence for Scores Derived from the Behavioural Regulation in Exercise Questionnaire

Katie E. Gunnell, Philip M. Wilson, Bruno D. Zumbo, Peter R.E. Crocker, Diane E. Mack, and Benjamin J.I. Schellenberg

Measurement is a process whereby a construct of interest is conceptually and operationally defined, and a scale is developed through which numbers are assigned to cases that imply the degree of that variable (Kline 2005). More simply stated, measurement answers the questions of 'why' and 'how' scores are assigned to variables (Wilson et al. 2011). Measurement is used to assess individuals for the purpose of research, intervention or feedback, decision-making, and potentially the creation of policy to enhance public welfare (Zumbo 2009). Because the preponderance of research in exercise psychology relies on the measurement of psychological or behavioral variables to answer complex questions, it is essential that scores and interpretations derived from these instruments demonstrate evidence of validity and reliability (Hagger and Chatzisarantis 2009; Zhu 2012). It is only then that evidence based on these scores or inferences can be fair, interpretable, relevant, and have functional worth in terms of social consequences (Messick 1998). Unfortunately, construct validity is all too often overshadowed by other agenda in the

K.E. Gunnell (✉) • P.R.E. Crocker
School of Kinesiology, The University of British Columbia, 210 War Memorial Gym, 6081 University Boulevard, Vancouver, BC V6T 1Z1, Canada
e-mail: kgunnell@interchange.ubc.ca; pcrocker@ubc.ca

P.M. Wilson • D.E. Mack
Behavioural Health Sciences Research Lab, Department of Kinesiology, Brock University, Niagara Region, 500 Glenridge Ave., St. Catharines, ON L2S 3A1, Canada
e-mail: phwilson@brocku.ca; dmack@brocku.ca

B.D. Zumbo, Ph.D.
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.J.I. Schellenberg
Department of Psychology, The University of Manitoba, Winnipeg, MB R3T 2N2, Canada
e-mail: schelle9@myumanitoba.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_10,

research process and remains an underappreciated aspect of exercise psychology research (Hagger and Chatzisarantis 2009).

The purpose of this paper is to examine how researchers who have used the Behavioural Regulation in Exercise Questionnaire (BREQ; Mullan et al. 1997) have applied validity theory to their investigations. In the following sections, a brief overview of the nature of validity is provided along with a synopsis of instrument development research culminating in the BREQ. The next section presents a systematic review of validity practices and results using the BREQ. The systematic review is designed to showcase how aspects of validity theory have been utilized (or omitted) in the development of a key instrument central to the study of motivation in exercise settings.

The Standards

The Standards for Educational and Psychological Testing (*The Standards*; American Educational Research Association [AERA], American Psychological Association [APA], and the National Council on Measurement in Education [NCME], 1999) serve as a guide to all researchers and applied practitioners by providing information about validity theory and validation practices[1]. Validity is defined as "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (AERA et al. 1999 p. 9). Prominent validity theorists (Kane 2001; Messick 1995) along with the current version of *The Standards* (AERA et al. 1999) now recognize that validity refers to the interpretation of test scores. Furthermore, there does not exist separate 'types' of validity. Validity theorists (Messick 1995) contend that score validation is an ongoing process in which various sources of evidence facilitate the interpretation and meaning of test scores (Sireci 2009). *The Standards* conceptualize validity as a unified concept, meaning that it is the degree to which all the evidence supports the intended interpretation of test scores derived from a given sample and context (AERA et al. 1999; Sireci 2009)[2]. Finally, score reliability should be included in the accumulation of validity evidence (AERA et al. 1999) because the assessment of measurement error helps researchers interpret the data quality from which inferences will be made (Zumbo 2007).

[1] In addition to the Standards, there are other validity frameworks from which researchers could base their validation efforts. For example, Messick (1995) advocated a progressive matrix of validity based on construct validity. Kane (2001) calls for a program of evaluation that is based on an argument of score validity. Zumbo (2007) has developed a validity framework that focuses on sample homogeneity and context of validity. Finally, Borsboom et al. (2004) proposed a conceptual framework for validity that opposes the unified view. Researchers are encouraged to read the original documents for a more in depth understanding of each framework.

[2] *The Standards* were never intended to be a prescription, or cookbook type document in which researchers can follow (Goodwin 2002). The process of validation is ongoing and as such, is based on the accumulation of knowledge.

The current edition of *The Standards* list five sources of validity evidence based on (a) content, (b) internal structure, (c) response processes, (d) relations to other variables, and (e) consequences of test use (see AERA et al. 1999 and Goodwin 2002 for a review). Evidence based on content refers to the wording, themes and format of items in addition to guidelines for administering the instrument and scoring protocols (AERA et al. 1999). For example, 'experts' can review the representativeness of BREQ items in relation to the theoretical constructs the items are intended to assess (AERA et al. 1999). Evidence based on internal structure represents the extent to which items from an instrument conform to the hypothesized construct(s) (AERA et al. 1999). For example researchers could conduct confirmatory or exploratory factor analysis to examine if items comprising the BREQ support the theoretically-based measurement model. Evidence based on response processes refers to the extent to which the performances or responses of participants match the intended interpretation of the construct (Goodwin 2002). For example a researcher could analyze how individuals respond to BREQ items through think aloud investigations. Finally, evidence based on consequences of testing refers to the potential intended and unintended consequences of using a particular test (AERA et al. 1999). For example, a researcher could conduct a descriptive study about the degree to which anticipated benefits from testing are analyzed (Goodwin 2002). An example using the BREQ is provided in the discussion section.

The Standards also make finer distinctions among the five sources of evidence. For example, evidence based on relations to other variables also includes criterion concurrent, predictive and discriminant/convergent evidence. Criterion concurrent evidence is how well scores gathered with one instrument correlate with scores from another *criterion* instrument of the *same construct* administered at the same time (AERA et al. 1999; Goodwin 2002). Criterion predictive evidence refers to how well scores from an instrument predict criterion scores at a later date (AERA et al. 1999). Convergent and discriminant evidence are provided through relationships between instrument scores and other instruments designed to measure similar (convergent) or different (discriminant) constructs (AERA et al. 1999). Campbell and Fiske's (1959) Multitrait-multimethod (MTMM) approach provides an examination of both convergent and discriminant evidence.

Unfortunately, as Zhu (2012) noted in exercise and sport psychology, there appears to be a rift between modern validity theory and applied research. Recognizing this disconnect, narrative reviews have called into question validation practices in exercise and sport psychology research (Haggar and Chatzisarantis 2009; Zhu). Hagger and Chatzisarantis (2009) advocated for greater attention to select validity practices. More recently, Zhu provided a cogent argument based on *The Standards* (AERA et al. 1999) around validity and validation practices in sport and exercise psychology. Zhu argued that certain validation efforts in our field need to "catch up" and "improve" (p. 19). More specifically, Zhu critiques authors in exercise and sport psychology for using outdated nomenclature when referring to validity 'types' and using incorrect labels when describing the evidence reported (e.g., citing concurrent evidence when the authors actually report convergent

evidence). To date, no investigation has systematically examined how validity theory has been applied to the validation of scores from a particular instrument in exercise or sport psychology. Using previous narrative commentaries as a starting point, we argue that aspects of validity theory warrant further consideration in order to advance the quality of validation efforts for scores of the BREQ.

Behavioural Regulation in Exercise Questionnaire: A Brief Overview

The BREQ is a 15-item instrument designed within the framework of self-determination theory (SDT; Deci and Ryan 2002) to assess both autonomous and controlled reasons that motivate exercise participation (Mullan et al. 1997). The items comprising the BREQ were created initially by modifying items contained within the Academic Motivation Scale (Vallerand et al. 1992) and the Self-Regulation Questionnaire for Academic settings (Ryan and Connell 1989)[3]. Consistent with SDT, the BREQ was designed to measure the following sources of exercise motivation: (a) *external* (i.e., motivation emanating from outside the self), (b) *introjected* (i.e., self-imposed contingencies such as to avoid negative emotions) (c) *identified* (i.e., motivation for personal value or purpose) and (d) *Intrinsic* regulations (i.e., doing an activity for its own sake). Markland and Tobin (2004) subsequently developed the BREQ-2, adding 4 items to assess *amotivation* (i.e., lack of intention to act). In the BREQ and BREQ-2, items designed to assess *integrated* regulation for exercise (i.e., value of the activity is congruent with sense of self) were not included despite the salience of this construct within SDT (Deci and Ryan 2002). Although *integrated* regulation is conceptually and theoretically distinct from *identified* regulation, operationally it is difficult to separate the concepts (Markland 2010). Recognizing the theoretical and practical implications of *integrated* regulation as a motivational resource in exercise, Wilson and colleagues (2006) created 4 items to measure *integrated* regulation designed to fit with existing BREQ and BREQ-2 items. Stemming from this work, McLachlan et al. (2011) have also developed 4 alternative *integrated* items to fit within the BREQ-2.

The BREQ is a popular instrument in exercise psychology research examining motivational issues using SDT as a framework. Variations in motives assessed by the BREQ have been linked to different stages of change (or readiness) for exercise (Daley and Duda 2006), exercise behavior (Mullan et al. 1997), and select forms of well-being (McLachlan et al. 2011; Wilson et al. 2006). The popularity of this instrument is further underscored by the number of language translations BREQ items have undergone (e.g., Spanish, Greek, Estonian, Persian). Additional lines of research attest to the instrument's flexibility given that the initial pool of BREQ

[3] For a more thorough review of the BREQ please review Mullan et al. (1997), and Wilson (2012).

items have been adapted for use in different contexts (e.g., physical education settings; BREQ-PE; Hein and Hagger 2007) and employed in a wide range of populations across the lifespan spanning from youth to older adults (Hagger et al. 2009; Markland and Tobin 2004).

Study Purposes and Research Questions

This study extends previous narrative commentary (Hagger and Chatzisarantis 2009; Zhu 2012) by outlining the status of validity theory as applied to research that has used the BREQ. We examined the following research questions:

1. What perspectives of validity theory have been used in score validation of the BREQ?
2. What sources of validity evidence have (and have not) been presented?
3. Were the sources of evidence presented consistent with guidelines set forth by the current version of *The Standards* (AERA et al. 1999)?

To accomplish these study purposes, we conducted a systematic review of published research using the BREQ and its variants with an emphasis on evidence for score validity reported in each study. The BREQ was chosen because it has received extensive use in various investigations and across cultures, languages and contexts.

Methods

Sampling

A computerized literature search was conducted to identify all peer reviewed full journal articles and book chapters that used the BREQ or a form of the BREQ. Searches were conducted for articles published from 1997 to June of 2012 using SCOPUS, Academic Search Premier including ERIC, MEDLINE, PsychARTICLES, PsychINFO, and SPORTDiscus. Title, subject, and abstract search words included: "behavioural regulation in exercise questionnaire" and "BREQ". A citation search for "Markland, Mullan, Ingledew 1997" was also used in each database. The reference list provided on The Exercise Motivation Measurement Index website (Markland 2010) was also cross-referenced.

Inclusion/Exclusion Criteria

All studies were considered, regardless of publication language. Research studies published in a language other than English were translated to determine if they met the criteria for inclusion in this study using a translation tool (www.translate.google.com)[4]. Theses, dissertations, conference abstracts or presentations were excluded from the sample. Research studies located using the procedure outlined above ($n = 174$) were screened to determine if a form of the BREQ (BREQ, BREQ-2, BREQ-2r [with integrated regulations] and BREQ-PE) was used. Studies were excluded if they did not report using any form of the BREQ ($n = 20$), the original BREQ had been modified whereby the author did not describe the modification in sufficient detail (e.g., the author indicated the research used 7-items from the BREQ only but did not indicate which items, the author used certain subscales of the BREQ but dropped items without providing an explanation; $n = 8$), the publication was a narrative review ($n = 4$), and the manuscript was describing protocols/intervention guidelines associated with a larger study ($n = 4$).

The primary author coded the remaining studies ($n = 138$) to determine if score validity evidence was explicitly presented[5]. The primary author and the sixth author coded the studies that presented explicit evidence of score validity ($n = 29$), using a fixed coding scheme to reduce ambiguity in the coding process. The primary coder had training in validity theory via doctoral level coursework and a publication history in areas of validation and exercise psychology constructs. The second coder was a graduate student who had completed undergraduate and graduate coursework in research methods and statistics, including a course on psychological testing. Absolute percent agreement between coders was 100 % for all validity categories coded except validity perspective (97 % agreement), and validity conception (76 % agreement). Disagreement pertaining to how a variable should be coded between coders was discussed and resolved prior to data analysis. Investigations that reported multiple studies within one manuscript were separated and coded independently (e.g., Wilson et al. 2006a, b, c).

[4] Using a translation tool represents a limitation of the current study. For studies that required translation, it is possible that the quality of the translation impacted coding and interpretations. Consequently, caution is warranted when interpreting our results, especially for investigations that were published in a language other than English.

[5] We acknowledge that under the unified conception of validity, all research informs the evidence base informing score validity. Consequently a limitation of this study is that we narrowed our search to only articles in which the authors explicitly discuss evidence of score validity. Therefore, the validity evidence presented herein may not represent all validity evidence available for scores of the BREQ; however, we believe the sampling strategy will allow for stronger conclusions about the application of validity theory to the BREQ.

Coding Scheme

Various aspects of each published study were coded prior to subsequent analysis to address the research questions in this study[6]. Information coded was based on conclusions and statements made by the authors of each coded study, regardless of whether or not the authors made conclusions in line with *The Standards* (AERA et al. 1999) about what source of evidence they were presenting. First, basic study information was coded including (a) sample (e.g., sport/athletes, pre-university physical education students, exercisers etc.), (b) age (e.g., under 18 years, 18–49.99 years, or over the age of 50 years; based on the mean age reported), (c) gender (e.g., mixed males and females, males only, females only), (d) ethnicity (e.g., mixed, specific race, not specified), (e) disease status (e.g., non-clinical or clinical), and (f) sample size. Next, the form of the instrument used was coded (e.g., BREQ, BREQ-2, BREQ-2r or BREQ-PE) and the language of the instrument was recorded (e.g., English, Greek). Coded studies were classified as either "psychometric investigations" or "research using the BREQ". Validity information was coded (see Table 10.1) based on a framework extrapolated from the work of Cizek and colleagues (2008), *The Standards* (AERA et al. 1999) and Goodwin's (2002) review of *The Standards*.

Data Analysis

First, basic demographic information was summarized to examine how the BREQ has been used and in what samples. Second, the frequency of studies reporting each source of validity information (see Tables 10.2, 10.3 and 10.4) was calculated. The nature of the evidence presented within each coded study was compared against the validity framework outlined by *The Standards* (AERA et al. 1999; see Table 5).

Results

Overview of Studies Included in the Review

Of the 138 studies coded, 29 studies explicitly described sources of validity evidence. Examination of the descriptive statistics indicated that most coded studies used adults aged 18.00–49.99 years old ($n = 22$) followed by children/youth under the age of 18.00 ($n = 6$) and adults over the age of 50.00 years ($n = 1$). The majority of coded studies used a sample of mixed gender ($n = 28$), one study used females

[6] Coding form can be obtained from the primary author upon request.

Table 10.1 Validity information coded

Information coded	Coding options
1. Reported estimates of score reliability as evidence of score validity	(a) Yes (b) No
2. Validity perspective or validity theory used by the authors of the investigation	(a) *The standards* (AERA et al. 1999) (b) Messick (1995) (c) Unsure/not clear
3. Conceptualization of validity	(a) Characteristic of the test (b) Characteristics of the test score/inference/interpretation (c) Mixed language (d) Unsure/not clear
4. Sources of validity evidence presented	(a) Evidence presented as content was further broken down into: content, face or unsure/not clear (b) Evidence based on internal structure was broken down to: factor analysis, item interrelationships, invariance tests, other, and unsure/not clear. *Note*. Descriptions of simplex patterns were coded as "other" even if the authors did not directly refer to it as validity evidence (c) Relationships to other variables were categorized into: convergent, divergent, discriminant, criterion predictive, criterion concurrent, criterion group differences, generalizations, construct and nomological networks. In corroboration with Cizek and colleagues (2008) investigation, evidence based on construct validity and nomological networks were included in the relations to other variables section. Although the current view of validity theory assumes that all evidence of validity bears on construct validity, it was coded separately if the authors described it as an independent source and *in relation to another variable*. If authors described results of factor analysis as 'construct validity' it was coded under internal structure (d) Response processes were categorized into: analysis of individual responses by interview and examining similarities/differences in responses by distinct groups or investigations in which researchers collect, record and interpret data (e) Consequences were categorized into: benefits associated with test use, negative uses, other and unsure/not clear

only, and no study examined males only. Ethnicity was not frequently reported ($n = 25$) with three studies reporting a mixed ethnicity in their sample and one study reporting a specific ethnicity. Only one coded study used a clinical sample (obese adolescents), with the remaining studies using a non-clinical/asymptomatic group of individuals (i.e., did not report any clinical diagnoses amongst participants).

Table 10.2 Instrument form and language

Instrument form	Number of studies
BREQ	9
BREQ-2	13
BREQ-2r	5
BREQ-PE	2
Language	
English	15
Persian	2
Greek	1
Spanish	4
Estonian	2
Dutch	1
Turkish	1
Romanian	1
Korean	1
Multi-language	1

BREQ Behavioural Regulation in Exercise Questionnaire, *BREQ-2* BREQ + amotivation scale, *BREQ -2r* BREQ + amotivation and integrated regulation scale, *BREQ-PE* BREQ that has been adapted to physical education contexts

Table 10.3 Validity perspectives

Validity perspective	Number of studies
Unitary perspective	0
Standards/Messick	6
Unsure/not clear	23
Conception	
Characteristic of the test	18
Characteristic of test score/inference/ interpretation	5
Mixed language	1
Unsure/not clear	5

Finally, sample sizes ranged from 51 to 1,071 in total across coded studies. The most frequently used version of the instrument was the BREQ-2. English was the most commonly used language of the instrument (see Table 10.2). Over half the coded studies were psychometric investigations ($n = 17$) and the remaining ($n = 12$) were research papers that used the BREQ, yet provided evidence of score validity.

Research Question 1: What Perspectives of Validity Theory Have Been Used in Score Validation of the BREQ?

Very few of the coded studies using the BREQ cited Messick (see Table 10.3). Conversely, the majority of the coded studies did not explicitly situate their

Table 10.4 Sources of validity evidence

Source of evidence	Number of studies
Content	4
Internal structure	27
Factor analysis	22
Item interrelationships	0
Invariance	4
Other: simplex pattern	16
Relations to other variables	13
Convergent	2
Divergent	1
Criterion-predictive	3
Criterion-concurrent	2
Criterion-group differences	0
Generalizations	0
Discriminant	8
Nomological network	4
Construct validity	1
Response processes	0
Consequences	0

Note. Within each of the five sources of validity evidence (i.e., content, internal structure, relations to other variables, response processes, and consequences), one paper may have reported more than one sub-source of validity evidence (e.g., factor structure, item interrelationships, invariance, and other). Therefore, sub-sources of validity evidence (e.g., factor structure, item interrelationships, invariance, and other) may not sum to equal the number of studies reporting that sources of validity evidence (e.g., internal structure evidence)

research within a validity framework. No researchers cited *The Standards*. Less than one-quarter described validity as a characteristic of the test score/interpretations or inference while over half described validity as a characteristic of the instrument (see Table 10.3).

Research Question 2: What Sources of Evidence Have (or Have Not) Been Presented?

The sources of validity evidence reported were based on content, internal structure and relationships to other variables (see Table 10.4). For internal structure, the majority of authors reported factor analysis, followed by invariance, and simplex structures. For relationships with other variables, authors most frequently reported evidence of discriminant, nomological networks, criterion predictive, criterion concurrent, convergence, divergence, and construct validity. Evidence based on response processes and consequences of test scores were not directly examined or reported in any coded study.

Research Question 3: Were the Sources of Evidence Being Presented Consistent with Guidelines Set Forth by the Current Version of The Standards?

When examining evidence of score validity, many investigators of the coded studies performed factor analysis, test score invariance between groups, examined a simplex pattern (yet few used procedures advocated by Jöreskog (1970) for assessing simplex structures), content evidence, convergent, divergent, nomological networks and construct evidence in line with the definitions and conceptualizations offered within *The Standards* (AERA et al. 1999). Evidence of score validity was not always examined through methods outlined by *The Standards* (AERA et al. 1999). One example is an investigation that confused score validity and reliability by reporting Cronbach's (1951) alpha as an estimate of 'internal validity'. Criterion predictive, criterion concurrent and discriminant evidence of score validity were not examined through methods consistent with definitions outlined by *The Standards* (AERA).[7]

Discussion

The overall purpose of this study was to outline the status of validity theory and evidence with respect to scores of the BREQ. To accomplish this purpose, a systematic review of studies that had used the BREQ or a form of the BREQ was analyzed for validity evidence using current frameworks for understanding score validity (e.g., *The Standards;* AERA et al. 1999). Consistent with previous speculations (Zhu 2012), results indicated that the majority of researchers using the BREQ have not embraced guidelines outlined within *The Standards*. Results of this systematic review also indicated that validity evidence for scores of the BREQ appear limited to content, internal structure and relations to other variables. Researchers using the BREQ have yet to directly examine validity evidence based on response processes or consequences.

Validity Theory as Applied to the BREQ

Relatively few researchers interpreted or provided explicit evidence of score validity in each study. Of the 138 studies identified as having used the BREQ, only 29 directly described validity evidence in their investigations. A common

[7] A detailed table containing authors' claims and how they were discrepant from *The Standards* can be obtained from the first author upon request.

trend found while coding articles was that authors would describe the BREQ in their methods section with statements such as 'The BREQ was previously validated' and provide a citation of authors who had previously examined evidence of score validity. In fact, this trend has become so pervasive in various literatures that it appears that questions regarding the validity of scores are all too often taken for granted (Bagozzi 1981; Hagger and Chatzisarantis 2009). Calling attention to this quandary, Hagger and Chatzisarantis (2009) note that if the validity of the inferences or scores is not assessed for a particular study (or sample), the findings of the investigation could be called in to question. Although journal space restrictions may preclude a full discussion of validity evidence, researchers would do well to at the very least, provide a sentence or two indicating that they examined some form of score validity (e.g., through confirmatory factor analysis). One investigation need not examine all possible sources of validity evidence, but rather report what is meaningful for their investigation (see Gunnell et al. 2014, Chap. 8).

While many researchers are undoubtedly cognizant that validity is a property of test scores or inferences, more consistency in the use of language in line with *The Standards* (AERA et al. 1999) when reporting study findings is warranted and justified to advance the field. By describing validity as a characteristic of the test, it perpetuates the notion that an instrument is either valid (or not) irrespective of context, sample and use. In turn, researchers may assume they can freely use the instrument across different contexts, samples, or languages without any concern for score validity. In addition, researchers should be cognizant that validity 'types' (e.g., 'convergent validity') are no longer established within the current iteration of *The Standards*. There is a subtle distinction between 'validity types' and 'types of evidence'. Put simply, validity is a unified conception based on the accumulation of validity evidence, not a checklist approach to find different compartmentalized validity 'types' (Goodwin 2002). Caution should be taken by using the term validity *evidence* (e.g., convergent evidence) such that old myths regarding 'types' of validity are dispelled rather than propagated. Researchers must be more conscientious in presenting validity evidence using accurate and standardized terminology (Zhu 2012) in order to advance the quality of measurement related research.

Researchers have not examined convergent, discriminant, and criterion evidence through methods advocated within the current version of *The Standards* (AERA et al. 1999). Authors who labelled their evidence as criterion 'concurrent validity' were often assessing convergent evidence (see Zhu 1998 for a full discussion on the distinction). In the psychometric and measurement literature, convergent and discriminant evidence are traditionally defined and assessed through correlations with scores for constructs derived from one or more *alternative instruments* (Campbell and Fiske 1959). This approach is advocated within the current version of *The Standards*; however, emerging validation techniques based on confirmatory factor analysis (CFA) have begun to appear in the literature (Kline 2005; Fornell and Larcker 1981). A large portion of the coded studies analyzed in this review using the BREQ employed CFA techniques exclusively to assess convergent and discriminant evidence. Discriminant evidence that purports to distinguish latent factors theorized to represent related subscales of one instrument may be

misinterpreted to be an examination of internal structure. For example, an evaluation of internal structure could include determining if there is one single factor or many correlated, yet distinct factors (AERA et al. 1999). This example demonstrates how *The Standards* definition of evidence based on internal structure is similar to how many researchers using the BREQ have examined discriminant evidence. *The Standards* are currently in the process of undergoing revision and we are hopeful for clarity concerning the definition of discriminant evidence. For now, researchers should take care to select, define, and justify techniques for assessing score validity that are based on validity theory and supported by experts.

Missing Sources of Validity Evidence for the BREQ

Validation efforts for the BREQ have focused on samples that are somewhat limited in terms of their breadth and scope. Very few of the coded studies adopted a specific ethnicity, or used different age groups such as youth or elderly. Furthermore, only 1 coded study used the BREQ in a clinical sample where disease was present. It would be beneficial for future researchers to examine more diverse and unique samples (e.g., pregnant women) and report evidence of generalizability for scores derived from the BREQ in such cohorts. Messick (1995) articulated a cogent argument concerning the issue of generalizability and setting the boundaries of score meaning. Zhu (2012) called for researchers to conduct validation studies across multiple samples to ensure the generalizability of the findings. This represents an important next step for researchers.

Lending support for Marsh's (1998) concern that factor analysis may become viewed as validity itself (Zhu 2012), it seems apparent that researchers using the BREQ have focused almost exclusively on analyses of internal structure using factor analytic techniques. Assessments of internal structure, such as through CFA's are important and researchers using the BREQ should be commended for their extensive use of this technique when presenting validity evidence. However, an overreliance on factor analysis could lead to a relatively limited view of score validity (Goodwin 2002). Different sources of validity evidence may reveal unique sources of score validity, as well as expand the cumulative knowledge pertaining to the nature and possibly function of the focal construct (e.g., exercise motivation). It is therefore important for researchers to examine alternative sources of validity evidence beyond merely internal structure (such as response processes and/or social consequences) where possible in order to build a stronger validity argument.

Evidence based on response processes and consequences of test use as sources of validity evidence have yet to be forthcoming in research using the BREQ. Validity evidence of response processes examines if participants were answering items as intended, or if construct irrelevant factors that confound the interpretation of test scores were operating during applied testing (Messick 1995). Investigators that seek to examine response processes (a concept akin to the substantive aspect of validity described by Messick 1995) could employ interviews with participants to ascertain

how or why the participant answered the way he/she did to individual test items (Goodwin 2002). Think-aloud protocols would also serve as a useful source of evidence pertaining to response processes (Sireci 2009). In the case of the BREQ, evidence based on response processes could examine if participants were providing true evaluations of reasons for exercise that motivate participatory behaviour or if social desirability response bias tendencies are contaminating responses to test items. This may be particularly important with the BREQ, as participants may feel pressured to provide responses to certain test items in a manner that is not consistent with their true feelings or underlying experiences in order to avoid negative consequences (e.g., being stereotyped or categorized as 'unmotivated').

Consequences as a source of validity evidence concern the intended and actual consequences of test score use (AERA et al. 1999). Messick (1995) advocated that consequences are related to value implications, score meaning and construct labeling. For example, a researcher who has designed an exercise intervention to increase intrinsic motivation based on existing exercise motivation could administer the BREQ in an attempt to classify individuals by their existing type of exercise motivation. If the scores from the instrument were biased (e.g., against women), it is possible that scores from the BREQ could incorrectly classify women's motivation, and in turn women may receive an intervention that does not match their existing exercise motivation. Therefore, in this example, the consequence of using scores of the BREQ to classify motivation can have potentially negative implications towards providing the correct intervention for women. Researchers using the BREQ should attempt to gather information on the consequences of score use, especially since scores from the BREQ could be used in applied testing scenarios or interventions.

Summary

Authors of the studies reviewed should be acknowledged for conducting validation studies; some have worked within modern validity frameworks, and some have conducted multi-study investigations or cross-validations to avoid the troublesome 'one shot' validity studies that permeate the literature (Zhu 2012). Many researchers using the BREQ have employed sophisticated data analysis procedures and examined evidence of score validity beyond the confines of the original validation efforts for the BREQ and their results generally support the inferences made from the scores of the BREQ. Notwithstanding these findings, researchers have generally not embraced contemporary validity theory and validation techniques outlined by *The Standards* (AERA et al. 1999). As such, further development and education regarding validity theory and validation practices are needed in this dynamic and growing field. Researchers are encouraged to maintain programs of research based on theoretical associations, and incorporate modern validity theory and validation procedures into the design and execution of their studies. Such approaches to research would be well grounded in validity frameworks that represent contemporary thinking and recommendations for practice within the

measurement literature (Zhu 2012), and as a consequence, advance the cumulative knowledge amassed within the field of exercise psychology with greater credence.

Acknowledgement The first and sixth authors were supported by a scholarship (doctoral award) from the Social Sciences and Humanities Research Council of Canada (SSHRC) during the preparation of this manuscript. The second and fifth authors were jointly supported by grant funding from the SSHRC during manuscript preparation and are affiliated with the Center for Bone and Muscle Health (Brock University). The fourth author was supported by a grant from SSHRC during manuscript preparation.

References[8]

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

Bagozzi, R. P. (1981). An examination of the validity of two models of attitude. *Multivariate Behavioral Research, 16*, 323–359. doi:10.1207/s15327906mbr1603_4.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071. doi:10.1037/0033-295X.111.4.1061.

Campbell, D. R., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. doi:10.1037/h0046016.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*, 397–410. doi:10.1177/0013164407310130.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

*Daley, A. J., & Duda, J. L. (2006). Self-determination, stage of readiness to change for exercise, and frequency of physical activity in young people. *European Journal of Sport Science, 6*, 231–243. doi:10.1080/17461390601012637.

Deci, E. L., & Ryan, R. M. (2002). Self-determination research: Reflections and future directions. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 431–441). Rochester: University of Rochester Press.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39–50.

Goodwin, L. D. (2002). Changing conceptions of measurement validity: An update on the new *Standards*. *Journal of Nursing Education, 41*, 100–106.

Gunnell, K. E., Schellenberg, B. J. I., Wilson, P. M., Crocker, P. R. E., Mack, D. E., & Zumbo, B. D. (2014). A review of validity evidence presented in the *Journal of Sport and Exercise Psychology* (2002–2012): Misconceptions and recommendations for validation research. In B. D. Zumbo & K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences*. Dordrecht: Springer.

Hagger, M. S., & Chatzisarantis, N. L. D. (2009). Assumptions in research in sport and exercise psychology. *Psychology of Sport and Exercise, 10*, 511–519. doi:10.1016/j.psychsport.2009.01.004.

*Hagger, M., Chatzisarantis, N. L. D., Hein, V., Soos, I., Karasai, I., Lintunen, T., & Leemans, S. (2009). Teacher, peer and parent autonomy support in physical education and leisure-time

[8] References marked with an asterisk indicate studies included in the validity evidence review. For a full list please contact the first author.

physical activity: A trans-contextual model of motivation in four nations. *Psychology & Health, 24*, 689–711. doi:10.1080/08870440801956192.

*Hein, V., & Hagger, M. S. (2007). Global self-esteem, goal achievement orientations, and self-determined behavioural regulations in a physical education setting. *Journal of Sports Sciences, 25*, 149–159. doi:10.1080/02640410600598315.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *The British Journal of Mathematical and Statistical Psychology, 23*, 121–145.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: The Guilford Press.

Marsh, H. W. (1998). Foreward. In J. L. Duda (Ed.), *Advances in sport and exercise psychology measurement* (pp. xv–xix). Morgantown: Fitness Information Technology.

Markland, D. (2010). *Exercise motivation measurement index*. Retrieved from http://www.bangor.ac.uk/~pes004/exercise_motivation/scales.htm

*Markland, D., & Tobin, V. (2004). A modification to the behavioural regulation in exercise questionnaire to include an assessment of amotivation. *Journal of Sport & Exercise Psychology, 26*, 191–196.

*McLachlan, S., Spray, C., & Hagger, M. S. (2011). The development of a scale measuring integrated regulation in exercise. *British Journal of Health Psychology, 16*, 722–743. doi:10.1348/2044-8287.002009.

Messick, S. (1995). Validity of psychological assessment: Validations of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–774.

Messick, S. (1998). *Consequences of testing interpretation and use: The fusion of validity and values in psychological assessment (Research Report)*. Princeton: Educational Testing Service.

*Mullan, E., Markland, D., & Ingledew, D. K. (1997). A graded conceptualization of self-determination in the regulation of exercise behaviour: Development of a measure using confirmatory factor analytic procedures. *Personality and Individual Differences, 23*, 745–752. doi:10.1016/s0191-8869(97)00107-4.

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalisation: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology, 57*, 749–761. doi:10.1037/0022-3514.57.5.749.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte: Information Age Publishing.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Brière, N. M., Senécal, C., & Vallières, E. F. (1992). The academic motivation scale: A measure of internal, external and amotivation in education. *Educational and Psychological Measurement, 52*, 1003–1017. doi:10.1177/0013164492052004025.

Wilson, P. M. (2012). Exercise motivation. In G. Tenenbaum, R. C. Eklund, & A. Kamata (Eds.), *Measurement in sport and exercise psychology* (pp. 293–302). Champaign: Human Kinetics.

*Wilson, P. M., Rodgers, W. M., Loitz, C. C., & Scime, G. (2006). It's who I am . . . Really! The importance of integrated regulation in exercise contexts. *Journal of Applied Biobehavioral Research, 11*, 79–104. doi:10.1111/j.1751-9861.2006.tb00021.x.

Wilson, P. M., Mack, D. E., & Sylvester, B. D. (2011). When a little myth goes a long way: The use (or misuse) of cut-points, interpretations and discourse with coefficient-alpha in exercise psychology. In A. M. Columbus (Ed.), *Advances in psychology research* (Vol. 77). Hauppauge: Nova.

Zhu, W. (1998). Comments on "development of a cadence curl-up test for college students" (Sparling, Millard-Stafford, & Snow, 1997): Concerns about validity and practicality. *Research Quarterly for Exercise and Sport, 69*, 308–310.

Zhu, W. (2012). Measurement practice in sport and exercise psychology: A historical comparative, and psychometric view. In G. Tenenbaum, R. C. Eklund, & A. Kamata (Eds.), *Measurement in sport and exercise psychology* (pp. 293–302). Champaign: Human Kinetics.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Pyschometrics, Vol. 26, pp. 45–79). Amsterdam: Elsevier Science.

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*. Charlotte: Information Age Publishing.

Chapter 11
Synthesis of Validation Practices in Two Assessment Journals: *Psychological Assessment* and the *European Journal of Psychological Assessment*

Anita M. Hubley, Sophie Ma Zhu, Ayumi Sasaki, and Anne M. Gadermann

Validity lies at the foundation of measurement and testing, as "without validity, a test, measure, or observation and any inferences made from it are meaningless" (Hubley and Zumbo 1996, p. 207). The *Standards for Educational and Psychological Testing*[1] (American Educational Research Association (AERA) et al. 1999) describe validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" and validation as a process that "involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (p. 9). They list five sources of evidence, namely evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. Though the fundamental importance of validity and validation is widely acknowledged, and the importance of investigating and reporting relevant evidence when using measurement instruments is strongly advocated by influential sources (e.g., AERA et al. 1999; Wilkinson and The APA Task Force on Statistical Inference 1999), previous test validation synthesis studies that examined the reporting practices of validity evidence of instrument developers and users indicate that the information provided in published articles is often insufficient and lacking (e.g., Cizek et al. 2008; Hogan and Agnello 2004; Qualls and Moss 1996).

Although numerous studies have examined validation practices in published research, there are no clearly agreed upon methods for conducting such test validation syntheses. Many systematic reviews of applied validation research focus on all articles, or a random or systematic selection of articles, that have

[1] To be referred to henceforth as *The Standards*.

A.M. Hubley (✉) • S.M. Zhu • A. Sasaki
Department of Educational and Counseling Psychology, and Special Education,
The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
e-mail: anita.hubley@ubc.ca

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

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_11,

been published in an area or topic (with a variety of criteria for inclusion) within a specified time frame and review and summarize specific findings often with respect to a particular measure or group of measures (e.g., methods and quality of factor analysis used with the SF-36; de Vet et al. 2005; sensitivity and specificity rates of the Reliable Digit Span; Schroeder et al. 2012; analysis and reporting of predictive validity in violence risk assessment measures; Singh et al. 2013; reliability and criterion-related validity results for physical activity questionnaires; Helmerhorst et al. 2012). Other reviews focus on all articles that have been published in an area or topic (with a variety of criteria for inclusion) and conduct meta-analyses (e.g., of the validity of individual Rorschach variables; Mihura et al. 2013; of the validity of activity monitors; Van Remoortel et al. 2012; of predictive validity of prodromal criteria in schizophrenia; Chuma and Mahadun 2011).

Most test validation syntheses have focused on the frequency, as well as the type, of reliability and validity evidence reported for measurement instruments (e.g., Barry et al. 2014; Cizek et al. 2008; Hogan and Agnello 2004; Jonson and Plake 1998; Slaney et al. 2009, 2010; Qualls and Moss 1996). Articles included in these validation syntheses were typically empirical or research articles published in various psychology and health journals as well as entries from other sources, such as APA's *Directory of Unpublished Experimental Mental Measures* or the *Mental Measurements Yearbook*. A review of these syntheses indicate that (a) the frequency of reporting reliability and validity evidence seems to have increased generally over time (although this may vary by journal or field of study), (b) internal consistency estimates of reliability (usually Cronbach's alpha) are reported far more frequently than test-retest reliability estimates, (c) validity evidence is often not reported for all measures in a study, (d) some forms of validity evidence (e.g., evidence related to content, convergent evidence, factor structure) tend to be reported most often (although which ones seems to vary by journal or field of study), (e) validity evidence such as developmental changes, effect of experimental variables, response processes, and consequences of testing is rarely reported, (f) the amount of validity evidence provided in any given study or article tends to be limited and typically is poorly reported, (g) sample characteristics are often not taken into account when reporting reliability and validity evidence based on previous research, and (h) there is a continuing disconnection between validity theory, available test standards, and validation practice.

We were interested in further examining validation practices in two premier assessment journals by explicitly focusing on articles that provide validation evidence. Specifically, the objective of our study was to conduct a research synthesis of validation practices in the two assessment journals, *Psychological Assessment* (*PA*) and the *European Journal of Psychological Assessment* (*EJPA*). *PA*, published by APA, has a focus on empirical research on measurement and evaluation in the area of clinical psychology. As highlighted by Green et al. (2011) "as the premier assessment journal for APA, *PA* should be an exemplar of good psychometric reporting practices for all APA journals in which psychological measures are used" (p. 657). *EJPA*, a journal of the European Association of Psychological Assessment, publishes manuscripts from all domains of psychological assessment, with a focus on studies that report on the development of new measures or the advancement of existing ones.

Both journals have an international readership, but there may be differences in the emphases of studies published, with one journal being housed in North America and the other one in Europe. Furthermore, given that the foci of the two journals are slightly different, with *PA* highlighting clinical psychology and *EJPA* covering psychology more broadly, different constructs may be included in publications, which may influence the sources of validity evidence reported as well as the specific types of evidence and validation approaches used.

In the present study, we randomly selected 50 articles published in 2011 or 2012 in the journals *PA* and *EJPA* to examine and provide a comparison of recent validation practices and the presentation of such information in these two premier psychological assessment journals. Our focus was on reporting information about the sources of validity evidence as described in *The Standards* (AERA et al. 1999) that were included in studies in these articles and not on assessing the quality of the validity evidence provided.

Method

Article Sampling

The title and abstracts of articles from each of 2 years (2011, 2012) of the journals *PA* and *EJPA* were previewed to determine if they were eligible for inclusion in the study. Articles were deemed eligible if they appeared to provide validation evidence, as described in *The Standards* (AERA et al. 1999), for one or more measures. Across the years 2011 and 2012, *PA* and *EJPA* published a total of 199 and 81 articles, respectively. Of these, 139 *PA* articles and 60 *EJPA* articles were deemed eligible for the study. We randomly selected 25 % of the eligible articles for each journal; this resulted in a sample of 35 *PA* articles and 15 *EJPA* articles.[2] Once coding began, two articles were subsequently deemed ineligible from *PA* and two other articles were randomly selected to replace them. If an article consisted of more than one study, the studies were coded separately; this resulted in 39 *PA* studies and 18 *EJPA* studies being coded.

Coding

To provide some context for the validity evidence, we coded for translation/adaptation methods used, if relevant, and reliability of scores on the measure of interest. Translation/adaptation methods were included because a significant

[2] A reference list of the 50 articles randomly selected from *PA* and *EJPA* may be requested from the first author.

number of validation articles involved the use of translated/adapted measures and the documentation of procedures used in translation/adaptation may inform the validity evidence obtained. We examined whether translation/adaptation guidelines were cited, the methodology used, whether the qualifications of the translators were provided, and whether pilot tests were conducted. Reliability was examined because it serves as a necessary, although not sufficient, condition for validity and is relevant to the interpretation of certain validity evidence. We only examined reliability estimates based on the data collected in the study as this is what is most relevant to validation (Hubley and Zumbo 2013; Vacha-Haase et al. 2002) and coded for the types of reliability estimate (e.g., internal consistency, test-retest, item-total correlation, inter-rater) provided.

Next, we coded for the five sources of validity evidence described in *The Standards* (AERA et al. 1999): (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing. These different sources are described in more detail below.

Evidence based on test content. Test content refers to "the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring" (AERA et al. 1999, p. 11). Evidence based on test content may be obtained during test development or later. Typically, such evidence relates to the generation of test items based on a literature review, prior existing measures of the construct, clinical/research experience of the author, feedback from subject matter experts (SMEs), or feedback from the target population (i.e., experiential experts (EEs)), and the examination or rating of elements of the measure by SMEs, EEs, or some other group.

Evidence based on response processes. As noted by *The Standards* (AERA et al. 1999), "Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (p. 12). An examination of response processes is not limited to the respondents; if a measure relies on observers, scorers, or judges to evaluate respondents' performance, the psychological or cognitive processes used by these individuals should be examined to determine if it is consistent with the intended interpretation of scores. Thus, attention was paid to evidence collected in studies that involved things such as probing responses to items (e.g., think-aloud protocols, cognitive interviewing), documenting or recording responses to items, recording the time needed to complete the measure of interest, and post-test questionnaires or interviews.

Evidence based on internal structure. *The Standards* (AERA et al. 1999) identified evidence based on internal structure as "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al. 1999, p. 13). That is, theory or the conceptual framework for a measure may suggest that a single dimension or multiple dimensions should be present and that the latter might involve independent, related, or hierarchical dimensions. We identified the type of analysis conducted (e.g., exploratory factor analysis (EFA) – including principal

component analysis, confirmatory factor analysis (CFA), measurement invariance – including structural equation modeling, differential item functioning, and other approaches).

Evidence based on relations to other variables. This is a very broad area of evidence that includes an examination of (a) the relationship of scores on the measure of interest to scores obtained on measures of the same or similar (convergent) constructs and different (discriminant) constructs, (b) the relationship of scores on the measure of interest to scores on a criterion or criteria (test-criterion relationships), and (c) group differences based on theory or evidence that such a difference should be present or not. Validity generalization studies are also included under this source of validity evidence. In each case, one is determining the degree to which a given interpretation of scores on the measure of interest is supported by obtained results. Generally, a criterion refers to an outcome that is of primary interest wherein the measure of interest may be viewed as a short-cut to obtaining the information provided by a criterion because the latter is normally too expensive or time-consuming to obtain. This validation approach is different from an examination of the relationship between a predictor variable and an outcome variable that may be seen in other types of research studies. Thus, studies in which the researchers were interested in whether scores on a measure of interest of one construct predict behavior or performance of an entirely different construct were not included in this synthesis, even if they self-identified as validation studies. We also recorded the terms that researchers used to describe the validation process (e.g., convergent validity, concurrent validity) and whether these appeared to be used correctly.

Evidence based on consequences of testing. The consequences (i.e., intended social consequences and unintended side effects; Hubley and Zumbo 2011, 2013) of legitimate test score interpretation are "relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components" (AERA et al. 1999, p. 16). Moreover, when claims are made about the benefits of testing that extend beyond the direct interpretation of test scores (e.g., use of scores from a measure will result in reduced health costs), evidence is also needed. To assist in identifying evidence based on the consequences of testing, we looked for the use of words such as "consequences", "consequential validity", "effects of", "impact of", "implications", and "clinical implications".

We only coded evidence based on primary data used in a study; we did not code evidence cited from prior studies in the literature. We also coded evidence as it fit the above sources of validity (or types of evidence within a source) rather than as the authors might have identified the evidence. That is, if authors claimed to provide 'criterion evidence' but it fit the definition of 'convergent evidence', we coded it as convergent evidence. Finally, all relevant evidence was coded if it met the definitions above, even if the authors did not identify it explicitly as validity evidence.

Results

As noted earlier, we examined 39 studies (35 articles) in *PA* and 18 studies (15 articles) in *EJPA*. Both samples of articles included authors from a wide range of countries (*PA* = 17; *EJPA* = 14), although only six countries were in common between the two journals. Articles tended to be multi-authored (*PA*: 2–12 authors, M = 4.89, SD = 2.34; *EJPA*: 1–10 authors, M = 3.87, SD = 2.53), with most papers having authors from one or two countries (*PA*: M = 1.40, SD = 0.78; *EJPA*: M = 1.40, SD = 0.51).

In terms of the randomly selected sample of articles,[3] studies from *PA* were predominantly about English-language measures (74.4 %), followed by measures in German (7.7 %), French (5.1 %), and then other languages. The English-language studies were primarily with U.S. samples (75.8 %) but also came from the U.K., Australia, and Canada. Studies from *EJPA* were more mixed, with German-language measures leading (22.2 %), followed by measures in English (16.7 %), Spanish (16.7 %), Portuguese (11.1 %), and then other languages. The small number of studies that used English-language measures in *EJPA* was either collected in the U.S. or online. The samples used in *PA* studies were mostly college/university students (25.6 %), clinical or clinical/nonclinical mixed samples (20.5 %), inmates (15.4 %), or from the general community (15.4 %), with the rest consisting of specific samples (e.g., teachers, doctors, soldiers) (35.9 %). The samples used in *EJPA* studies were mostly college/university students (38.9 %) or from the general community (22.2 %), with the rest consisting of specific samples (50.0 %). Finally, in examining the kinds of measures used, studies from *PA* used predominantly clinical (e.g., depression, eating disorders; 48.7 %), forensic (e.g., violence risk appraisal, psychopathy; 23.1 %), or personality measures (7.7 %) whereas studies from *EJPA* used predominantly personality (e.g., sensation seeking, self-esteem; 38.9 %), clinical (27.8 %), positive psychology (e.g., well-being, character strengths; 11.1 %), or educational (e.g., schoolwork engagement; 11.1 %) measures.

Use of Translation

As seen in Table 11.1, the samples from the two journals differed in terms of the use of translation. The *EJPA* sample included a larger number of new translations (half of the studies) followed by use of previously translated measures and non-English measures with no translation (both 20.0 %) whereas the *PA* sample tended to include English measures with no translation (67.6 % of the studies) followed by

[3] Percentages in this section do not sum to 100 % as some studies may have included multiple language versions of a measure, used more than one sample, or collected data from more than one country.

Table 11.1 Use of translation in *Psychological Assessment* and *European Journal of Psychological Assessment* studies

Journal	Relevant samples/ studies[a]	No translation (English)	No translation (non-English)	New translation	Previously translated	Unclear when translated
PA	43[b]	29 (67.4 %)	1 (2.3 %)	3 (7.0 %)	6 (14.0 %)	4 (9.3 %)
EJPA	20	2 (10.0 %)	4[c] (20.0 %)	10[c] (50.0 %)	4 (20.0 %)	0 (0.0 %)
Total	**63**	**31 (49.2 %)**	**5 (7.9 %)**	**13 (20.6 %)**	**10 (15.9 %)**	**4 (6.3 %)**

PA Psychological Assessment, *EJPA* European Journal of Psychological Assessment

[a]There are more samples/studies included here than the 39 *PA* studies and the 18 *EJPA* studies because some studies used more than one language in a study

[b]One study used 'response latency' as a measure for which language is not relevant

[c]One study used an original measure in Turkish and also translated it into English

use of previously translated measures (14.0 %). Both journals reported use of measures that were administered in a variety of languages.

Of the three studies using newly translated measures in *PA*, only two studies (66.7 %) cited translation guidelines. Two studies (66.7 %) used forward and backward translation and one (33.3 %) used forward translation only. Generally, two individuals were involved in the translation process. Only two of the articles (66.7 %) provided any information about the translators' qualifications and that was limited to whether they were native or bilingual speakers. In *EJPA*, only three (30.0 %) of the ten studies using newly translated measures cited translation guidelines. Five studies (50.0 %) used forward and backward translation, one (10.0 %) used forward translation only, and four studies (40.0 %) provided no information about the translation process. Generally, multiple individuals were involved in the translation process. Only six articles (60.0 %) provided any information about the translators' qualifications and that was primarily limited to whether they were native or bilingual speakers. A pilot study was conducted in only one study (10.0 %), but very little information was provided.

Reliability

Reliability evidence was reported in similar percentages of studies from *PA* (89.7 %, 35 out of 39 studies) and *EJPA* (88.9 %, 16 out of 18 studies). Table 11.2 shows the frequency and percentage of different types of reliability evidence reported in the studies from each journal. Internal consistency was, in each case, by far the most commonly reported type of evidence, with Cronbach's alpha reported the vast majority of times (*PA*: 28 out of 31 studies, 90.3 %; *EJPA*: 15 out of 15 studies; 100 %). Studies in each journal reported, on average, 1.5 types of evidence. Two or more kinds of reliability evidence were presented in 28.2 % of cases (11 out of 39 studies) in *PA* and in 38.9 % of studies (7 out of 18 studies) in *EJPA*. Test-retest reliability correlations were the second most commonly reported evidence for *PA* whereas item-total correlations were the second most commonly reported evidence for *EJPA*.

Sources of Validity Evidence

Table 11.3 shows the frequency and percentage of studies in *PA* and *EJPA* providing each of the broad sources of validity evidence as outlined in *The Standards*. *PA* and *EJPA* studies were similar in their emphasis on internal structure (73.2 %) and relations to other variables (76.8 %) evidence, although a notably higher percentage of *EJPA* studies (83.3 %) provided internal structure evidence than did *PA* studies (68.4 %). Little to no evidence related to test content, response processes, or consequences of testing was found in the sample of studies examined

Table 11.2 Types of reliability evidence across *Psychological Assessment* and *European Journal of Psychological Assessment* studies

		Types of reliability evidence[a]						
Journal	Relevant studies	Internal consistency	Test-retest reliability	Inter-item correlations	Item-total correlations	IRT TIF/CSEM[b]	Inter-rater reliability	Number of reliability types
PA	35	31 (88.6 %)	7 (20.0 %)	4 (11.4 %)	5 (14.3 %)	3 (8.6 %)	3 (8.6 %)	M = 1.51 (SD = 0.92)
EJPA	16	15 (93.8 %)	2 (12.5 %)	2 (12.5 %)	5 (31.3 %)	0 (0.0 %)	0 (0.0 %)	M = 1.50 (SD = 0.63)
Totals	**51**	**46 (90.2 %)**	**9 (17.6 %)**	**6 (11.8 %)**	**10 (19.6 %)**	**3 (5.9 %)**	**3 (5.9 %)**	**M = 1.51 (SD = 0.83)**

PA Psychological Assessment, *EJPA* European Journal of Psychological Assessment

[a]Summing the frequencies and percentages will not equal the total number of relevant studies and 100 %, respectively, because some studies include more than one type of reliability evidence

[b]Item Response Theory (IRT) test information function (TIF) and conditional standard error of measurement (CSEM)

Table 11.3 Sources of validity evidence across *Psychological Assessment* and *European Journal of Psychological Assessment* studies

		Sources of validity evidence[a]					
Journal	Relevant studies	Test content	Internal structure	Relations to other variables	Response processes	Consequences of testing	Number of validation sources
PA	38[b]	0 (0.0 %)	26 (68.4 %)	29 (76.3 %)	1 (2.6 %)	0 (0.0 %)	M = 1.44 (SD = 0.55)
EJPA	18	0 (0.0 %)	15 (83.3 %)	14 (77.8 %)	0 (0.0 %)	0 (0.0 %)	M = 1.60 (SD = 0.50)
Totals	**56**	**0 (0.0 %)**	**41 (73.2 %)**	**43 (76.8 %)**	**1 (1.8 %)**	**0 (0.0 %)**	**M = 1.49 (SD = 0.54)**

PA Psychological Assessment, *EJPA* European Journal of Psychological Assessment

[a]Summing the frequencies and percentages will not equal the total number of relevant studies and 100 %, respectively, because some studies include more than one source of validity evidence

[b]One study in a multi-study article provided reliability evidence but did not provide validity evidence

Table 11.4 Types of internal structure evidence across *Psychological Assessment* and *European Journal of Psychological Assessment* studies

		Types of internal structure evidence[a]			
Journal	Relevant articles	EFA	CFA	Measurement invariance	Number of internal structure types
PA	26	14 (53.8 %)	19 (73.1 %)	12 (46.2 %)	M = 1.73 (SD = 0.72)
EJPA	15	6 (40.0 %)	9 (60.0 %)	4 (26.7 %)	M = 1.27 (SD = 0.46)
Totals	**41**	**20 (48.8 %)**	**28 (68.3 %)**	**16 (39.0 %)**	**M = 1.56 (SD = 0.67)**

PA Psychological Assessment, *EJPA* European Journal of Psychological Assessment, *EFA* exploratory factor analytic approaches, *CFA* confirmatory factor analytic approaches

[a]Summing the frequencies and percentages will not equal the total number of relevant studies and 100 %, respectively, because some studies include more than one type of internal structure evidence

here. Thus, we will examine internal structure and relations to other variables evidence in more detail.

Internal Structure. Table 11.4 shows the frequency and percentage of studies in *PA* and *EJPA* providing each of three types of internal structure evidence (i.e., EFA approaches, CFA approaches, and measurement invariance approaches). For both journals, CFA was presented most frequently (68.3 %), followed by EFA (48.8 %) and measurement invariance (39.0 %), but there were notable differences between the journals. In *PA*, 57.7 % of studies reported two or more types of internal structure evidence whereas only 26.7 % of *EJPA* studies did so. As a result, studies in *PA* reported a higher percentage of each type of evidence relative to studies in *EJPA*.

Relations to Other Variables. Table 11.5 shows the frequency and percentage of studies in *PA* and *EJPA* providing each of a variety of relations to other variables evidence. For both journals, convergent evidence was, by far, the most frequently reported evidence in studies at 83.7 %. Again, there were some notable differences between the journals. For studies in *PA*, the second most commonly reported evidence was incremental evidence at 41.4 %, followed by discriminant and criterion-related evidence; for studies in *EJPA*, the second most commonly reported evidence was criterion-related evidence at 21.4 %, followed by discriminant evidence. In *PA*, 65.5 % of studies reported two or more types of relations to other variables evidence whereas only 21.4 % of *EJPA* studies did so. Moreover, 24.1 % of *PA* studies reported three or more types of relations to other variables evidence whereas only 14.3 % of *EJPA* studies did so. Consequently, studies in *PA* reported a higher percentage of each type of evidence relative to studies in *EJPA*.

Table 11.5 Relations with other variables evidence across *Psychological Assessment* and *European Journal of Psychological Assessment* studies

		Relations with other variables evidence[a]						
Journal	Relevant studies	Convergent	Discriminant	Criterion	Known groups	Incremental	Treatment effects	Number of evidence types
PA	29	24 (82.8 %)	11 (37.9 %)	10 (34.5 %)	3 (10.3 %)	12 (41.4 %)	2 (6.9 %)	M=2.14 (SD=1.27)
EJPA	14	12 (85.7 %)	2 (14.3 %)	3 (21.4 %)	1 (7.1 %)	1 (7.1 %)	0 (0.0 %)	M=1.36 (SD=0.75)
Totals	**43**	**36 (83.7 %)**	**13 (30.2 %)**	**13 (30.2 %)**	**4 (9.3 %)**	**13 (30.2 %)**	**2 (4.7 %)**	**M=1.88 (SD=1.18)**

PA Psychological Assessment, *EJPA European Journal of Psychological Assessment*

[a]Summing the frequencies and percentages will not equal the total number of relevant studies and 100 %, respectively, because some studies include more than one type of relations with other variables evidence

Discussion

In this validation synthesis, we aimed to examine and provide a comparison of recent validation practices and the presentation of such information in the journals *PA* and *EJPA* based on a random selection of 50 articles published in 2011 or 2012. When examining our findings and comparing them to previous studies examining reliability or validation practices, it is important to remember that we selected: (a) two journals that would be considered premier journals in the areas of assessment, measurement, and validation, and (b) articles that appeared to provide validation evidence rather than simply empirical or research articles.

There are some notable similarities and differences between the samples of studies from *PA* and *EJPA*. Both samples of articles included authors from a wide range of countries, tended to be multi-authored, and tended to include authors from one or two countries. While authors in both journals came from a wide range of countries, only six countries were in common between the two journals. Studies from *PA* were predominantly about English-language measures used with mostly U.S. samples. The language of measures and nationalities of samples reported on in *EJPA* were mixed, with more being German, English, Spanish, and Portuguese. Studies in both journals relied heavily on college/university student and community samples but studies in *PA* also focused on clinical or clinical/nonclinical mixed samples and inmate samples. Studies in both journals tended to examine clinical measures, but *EJPA* also tended to focus on personality measures whereas *PA* studies tended to focus on forensic measures, which were not examined in any of the *EJPA* studies.

Translation/Adaptation

The samples from the two journals differed in terms of the use of translation/adaptation with the *EJPA* sample including a larger number of new or previously translated measures whereas the *PA* sample tended to include more English measures with no translation. Relatively few studies cited translation guidelines, although a greater percentage of articles in the *PA* sample did so. In some cases, citation of translation guidelines seemed to take the place of describing the translation/adaptation procedures used in a study, which leaves the reader uncertain about exactly what procedures were used. Generally, multiple individuals were involved in the translation process but only about two-thirds of articles provided any information about the translators' qualifications and that was primarily limited to whether they were native or bilingual speakers. Most studies used forward and backward translation, with far fewer studies using forward translation only; notably, however, 40 % of the studies in *EJPA* provided no information about the translation process. In addition, use of a pilot study was reported in only one study and very little information was provided. Our findings are more positive than those

of Whittington (1998), who found that only 17 % of articles described how the measure was developed or adapted and only 15 % reported whether it was piloted before final use. Nonetheless, it is clear that more attention needs to be paid to the translation or adaptation procedures used with measures and more detail needs to be provided when reporting those procedures. Specifically, authors need to clearly state whether measures used in their research are original measures in that language, have been previously translated or have been translated for the current study. Previously translated measures should be appropriately cited. Translation or adaptation methods used for measures in a current study need to be clearly presented with detailed information about the procedures used, the qualifications of translators, the use of guidelines, and details about any pilot work conducted or feedback obtained.

Reliability

Reliability evidence was reported in similarly high percentages of studies from *PA* (87.7 %) and *EJPA* (88.9 %). These percentages are notably higher than has been reported in other recently published studies. For example, Barry et al. (2014) found that only 42.3 % of their articles reported reliability estimates based on their own samples' responses in seven prominent journals that they considered representative of health education and health behavior. Green et al. (2011) examined reliability reporting practices in a sample of empirical articles using self-report measures from the 1989, 1996, and 2005 volumes of *PA* but found only 10–28 % of researchers reported reliability estimates based on their own samples' responses. Vacha-Haase et al. (2002) found that, in a meta-analysis of reliability generalization studies, few empirical studies (24.4 %) report reliability estimates based on the samples' scores. Similarly low percentages are reported by Vacha-Haase et al. (1999), Thompson and Snyder (1998), Qualls and Moss (1996), Meier and Davis (1990), and Willson (1980). Closer is Slaney et al.'s (2009) finding that 72.8 % of articles reported reliability estimates generated from sample responses in their study of empirical articles from four journals (which included *PA*) in 2004. Overall, our higher percentages likely reflect the journals and types of articles we selected; that is, we chose two journals that focus on assessment and measurement and we selected among articles that conducted validation work.

In terms of the reliability evidence provided, internal consistency was, by far, the most commonly reported type of evidence, appearing in 88.6 % of such studies in *PA* (90.3 % using Cronbach's alpha) and 93.8 % of such studies in *EJPA* (100 % using Cronbach's alpha). The second most common reliability estimates was test-retest reliability coefficients, which were found in 20.0 % of the reliability studies in *PA* and 12.5 % of these studies in *EJPA*. Our percentages of internal consistency reliability estimates are somewhat higher than percentages reported by others. For example, Barry et al. (2014) reported that the most frequent types of reliability estimates in their study were internal consistency estimates (74.2 %, with 88.6 %

being Cronbach's alpha). Hogan et al. (2000), in an examination of a sample of entries from one volume of the *Directory of Unpublished Experimental Mental Measures* covering tests appearing in journals from 1991 to 1995, found that approximately 77 % of entries reported internal consistency estimates (with approximately 86.5 % using Cronbach's alpha). Our percentages of studies reporting test-retest reliability coefficients are roughly in line with others as Barry et al. (2014) reported that 15.3 % of reliability estimates were test-retest reliability coefficients. Likewise, Hogan et al. (2000) found that 19 % of entries reported test-retest reliability coefficients. All other types of reliability estimates (e.g., inter-rater, parallel forms, test information function) are reported rarely, if at all, in both our analysis and in other studies.

Two or more kinds of reliability evidence were presented in 28.2 % of studies in *PA* and in 38.9 % of studies in *EJPA*. These percentages are much higher than those reported in previous literature. Multiple types of reliability evidence were provided in only 8.8 % of articles in Qualls and Moss (1996) and in less than 20 % of tests in Hogan et al. (2000). Together, internal consistency and test-retest reliability evidence appeared in only 5 % of studies in Barry et al. (2014). Test-retest reliability correlations were the second most commonly reported evidence for *PA* whereas item-total correlations were the second most commonly reported evidence for *EJPA* in our study.

Overall and like previous research, we found that there tends to be a lack of consideration of the language, sample/subsamples, and conditions under which measures are collected when relying on reliability evidence despite the fact that it is well known in the measurement field that reliability is a characteristic of the scores on a measure obtained from a sample or subsamples (Crocker and Algina 1986; Green et al. 2011; Helms et al. 2006; Traub 1994; Vacha-Haase et al. 2000; Whittington 1998; Willson 1980). Researchers need to pay more explicit attention to the sample providing evidence related to reliability of scores and validity of inferences before relying on that evidence for their own purposes.

Appropriate reliability estimates should be reported for the scores obtained for each measure used in a study and greater attention needs to be paid to estimates or information besides Cronbach's alpha. Estimates used in a study should be clearly identified (e.g., as Cronbach's alpha – not just internal consistency, inter-rater reliability using Cohen's kappa). When measures are comprised of multiple subscales, reliability estimates should be reported for the scores obtained on each subscale. Moreover, when different subgroups or samples (e.g., men/women, clinical/non-clinical groups, different ethnic/racial/nationality groups) are examined in a study or different language versions of measures are used, reliability estimates should be reported for scores obtained from measures for each subgroup/sample or language version.

Sources of Validity Evidence

Many validation synthesis studies have found that few studies provide validity evidence to support the inferences made from the measures used in their research (Barry et al. 2014; Hogan and Agnello 2004; Qualls and Moss 1996; Slaney et al. 2009, 2010; Whittington 1998). For example, Qualls and Moss (1996) reported that validity information was provided for only 31.7 % of measures examined (with over one-third of this evidence based on previous research) in their study of articles published in 14 APA journals in 1992 and Barry et al. (2014) found that only 26 % of articles published between 2007 and 2010 in seven journals in the area of health education and health behavior reported validity evidence. Even Hogan and Agnello (2004) reported that only 54.6 % of their sample of entries from APA's *Directory of Unpublished Experimental Mental Measures*, covering tests appearing in journals from 1991 to 1995, reported some type of validity evidence. Slaney et al. (2009), however, reported that validity evidence was generated from the study sample in 92.4 % of articles appearing in the journals *Educational and Psychological Measurement*, *Psychological Assessment*, *Journal of Personality Assessment*, and *Personality and Individual Differences* in 2004 but these are all journals that focus on measurement, assessment, and validation issues. Likewise, in our study, we focused on articles that provided validity evidence, so nearly all of the studies provided validity evidence and any exceptions were reliability-only studies from multi-study papers.

Similar to some previous research (Barry et al. 2014; Cizek et al. 2008, 2010; Hogan and Agnello 2004; Qualls and Moss 1996; Slaney et al. 2009), our sample of studies from *PA* and *EJPA* show that internal structure and relations to other variables sources of validity evidence are strongly favored, with little to no evidence presented related to test content, response processes, or consequences of testing. There was some evidence that the emphasis might differ from journal to journal given that a higher percentage of *EJPA* studies provided internal structure evidence relative to *PA* studies. The fact that these journals focused more on clinical, forensic, or personality tests might explain the dearth of validation evidence related to content in this sample of studies. Cizek et al. (2008) found that evidence related to test content tended to be presented more frequently in achievement, developmental, and motor skills tests and less frequently in personality/psychological, attitude, and vocational tests and Whittington (1998) found content validity was reported for 45 % of measures in educational research. One consideration that arises with the sources of response processes or consequences of testing is that, relative to the other sources of validity evidence, there is less clear and accepted practice about how to obtain such evidence or present it. This makes such evidence harder to locate in the literature. Moreover, some potential sources of response processes seem to blur with evidence that might be presented as part of test content.

We subsequently examined the kinds of evidence that were presented under internal structure and relations to other variables sources of validity evidence.

In terms of evidence related to internal structure, CFA was presented most frequently (68 %), followed by EFA (49 %) and measurement invariance (39 %). We are not aware of previous literature that has examined the types of internal structure evidence examined. In this volume's chapter by Chinni and Hubley, which examined validation practices with the Satisfaction with Life Scale from 1985 to 2012, about 85 % of studies examined internal structure, with 59 % using CFA and only 31 % using EFA.

With respect to relations to other variables evidence, convergent evidence is, by far, the most frequently reported at 84 % in both *PA* and *EJPA*. Generally, criterion-related and discriminant evidence were the next most commonly reported kinds of evidence but it is noteworthy that 41 % of *PA* studies also reported incremental evidence (whether they used this exact term or not). For example, a researcher might indicate finding that scores on a measure predict criterion scores beyond that provided by other variables which supports the utility of a measure or supports validity. While this type of evidence might have been subsumed under criterion-related evidence in some previous validation synthesis articles, we coded for it separately and found that it was the second most commonly reported evidence in our sample of *PA* studies but was rarely reported (<10 %) in our sample of *EJPA* studies.

Few studies appear to have examined the frequency of reporting multiple types of relations to other variables evidence. In *PA*, 65.5 % of studies reported two or more types of relations to other variables evidence whereas only 21.4 % of *EJPA* studies did so. These percentages are much higher than those reported in the literature, which tend to be about 5 % or less (Hogan and Agnello 2004; Qualls and Moss 1996).

There are a number of issues that arise when examining relations with other variable evidence. One issue has to do with the high frequency of convergent evidence but relatively low inclusion of discriminant evidence. While it is not surprising that convergent evidence is presented more frequently than discriminant evidence, it is disappointing. When conducting a study using convergent measures, it is important to: (a) include discriminant measures for comparison, and (b) specify in advance the expected relative magnitude of coefficients from, or the rank order of, each of the convergent and discriminant measures. As Hubley and Zumbo (2013) noted, it is useful to think of convergent and discriminant measures as being on a continuum wherein correlations between theoretically similar measures (i.e., convergent validity) should be 'relatively high' while correlations between theoretically dissimilar measures (i.e., discriminant validity) should be 'relatively low'. This permits the researcher to better interpret the obtained validity coefficients.

In many cases, there appears to be little or no intention to include discriminant measures in a study, but low correlations found between scores on the measure of interest and other measures are subsequently described as 'discriminant' in discussion sections. The researcher should not be presenting a table of validity coefficients and then label evidence as convergent or discriminant based simply on the significance, direction, or magnitude of the obtained coefficients. There is also a

tendency for researchers to focus on supporting evidence and ignore evidence (such as low convergent coefficients) that does not support the intended inferences. This is most evident when a table of correlations among variables is presented but only some correlations are discussed. The researcher should explain exactly why specific convergent and discriminant measures are being selected and specify, in advance, the evidence that is needed to support validity claims about the intended inferences from measures. Otherwise, how will the researcher (and reader) be able to judge the degree to which the evidence supports (or not) the intended inferences?

Another issue is that while the terms 'concurrent' and 'predictive' have been associated with criterion-related evidence in theoretical writings in validity and validation, these terms (and especially 'concurrent') are frequently used with convergent evidence. Moreover, if convergent evidence is examined using regression techniques, the term 'predictive' may be used even if the variables are collected at the same point in time. Another problem is that the phrase 'predictive validity' may be used but does not always refer to validity evidence per se but sometimes simply to the use of statistical prediction. But this also raises questions about what qualifies as a 'criterion' in criterion-related evidence. There are many cases in which the evidence being presented should be treated as convergent evidence rather than criterion-related evidence. An example of this would be examining the correlation between scores on two depression screens; even if one screen might be considered a 'gold standard' in the field, it is better conceptualized as a convergent measure than as a criterion. There are also cases, however, in which one needs to decide whether the so-called criterion is truly an outcome variable as envisioned in validation work or whether it is simply a dependent variable in a research study. Essentially, how closely related to the construct represented by the measure of interest does the so-called criterion need to be? One needs to consider whether the ability of scores on the measure of interest to explain or predict an outcome/dependent variable (i.e., so-called criterion) contributes to our understanding of the inference being made from scores on the measure of interest. In some cases, researchers attempt to do too much in one study and try to both provide validity evidence (i.e., conduct a validation study) and determine the contribution of a predictor to the dependent variable (i.e., conduct a research study).

Yet another issue is that researchers do not seem to have a handle on how to examine or present known-groups validity evidence. In known-groups validation, the researcher needs to present theoretical and/or empirical evidence of a known difference among the groups on the construct of interest. Consequently, if a significant difference is then shown to exist among the groups, then evidence is present to support the intended inference from the scores on that measure. At the same time, if such a difference is not found, then the researcher must acknowledge that the evidence does not support the intended inference. It is crucial then to select one's groups carefully. Even when researchers intend to provide known-groups evidence, their interpretation and reporting of this evidence tends to be weak. That is, there is a tendency for evidence that supports the intended inference to be identified as supporting validity but evidence that does not is simply discussed descriptively. At the same time, researchers too often seem to confuse exploring

possible group differences (e.g., sex differences) with known-groups validity evidence.

In many cases, researchers appear to have far more evidence at hand than they actually present. This is particularly the case for discriminant evidence and for known-groups evidence. In the case of discriminant evidence, for example, researchers may not always consider that not all subscales on measures used in the study would be expected to be correlated and thus miss the opportunity to provide discriminant evidence and, at the same time, strengthen their presentation of convergent evidence. In the case of known-groups evidence, researchers sometimes simply explore group differences in their studies without recognizing that a well-known difference could be presented as known-groups evidence.

Finally, known-groups evidence is sometimes described as discriminant evidence with the idea that the measure can be used to discriminate group membership. This may be dependent on the field of study as Hubley (2014) has pointed out that this mislabeling of discriminant validity evidence is extremely common in the quality of life literature. On the one hand, it is tempting to argue that the use of terms to describe different types of relations with other variables evidence is an unnecessary distraction given that they all qualify as validity evidence. This is certainly true when one focuses on the degree to which there is a preponderance of evidence to support the intended inferences from scores on the measure of interest. On the other hand, the correct identification of these types of evidence is important if we are to learn more about the types of evidence that are favored in validation work, understand where gaps in such evidence exist, and provide better education about how such evidence might be presented and interpreted.

Limitations

It is important to recognize the limitations that might exist in a validation synthesis study. The small number of studies (25 %) sampled from *PA* and *EJPA* in this study may not fully represent validation practices conducted by all studies reported within these journals or by studies within the field as a whole. In addition, in coding validation procedures we were constrained by the clarity, and quality, of the information reported in studies and, at times, we were required to use our best judgment to evaluate the presence of certain types of evidence. Nonetheless, our findings do seem to be fairly consistent with the findings of previous validation synthesis research. It is also worth noting that our results, like those in other validation syntheses, may not strictly reflect what researchers know or do with respect to validation practices but reflect other factors such as editorial policies or restrictions (e.g., page length) or reviewer recommendations.

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.

Barry, A. E., Chaney, B. H., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: a review of seven journals. *Health Education & Behavior, 41*, 12–18.

Chuma, J., & Mahadun, P. (2011). Predicting the development of schizophrenia in high-risk populations: Systematic review of the predictive validity of prodromal criteria. *The British Journal of Psychiatry, 199*, 361–366. doi:10.1192/bjp.bp.110.086868.

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. doi:10.1177/0013164407310130.

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. doi:10.1177/0013164410379323.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

de Vet, H. W., Adèr, H. J., Terwee, C. B., & Pouwer, F. (2005). Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36.*Quality of Life Research, 14*, 1203–1218. doi:10.1007/s11136-004-5742-3.

Green, C. E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in psychological assessment: Recognizing the people behind the data. *Psychological Assessment, 23*, 656–669. doi:10.1037/a0023089.

Helmerhorst, H. F., Brage, S., Warren, J., Besson, H., & Ekelund, U. (2012). A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. *The International Journal of Behavioral Nutrition and Physical Activity, 9*, 103. doi:10.1186/1479-5868-9-103.

Helms, J. E., Henze, K. T., Sass, T. L., & Mifsud, V. A. (2006). Treating Cronbach's alpha reliability coefficients as data in counseling research. *The Counseling Psychologist, 34*, 630–660. doi:10.1177/0011000006288308.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*, 802–812. doi:10.1177/0013164404264120.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*, 523–531. doi:10.1177/00131640021970691.

Hubley, A. M. (2014). Discriminant validity. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 1664–1667). Dordrecht: Springer.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *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. doi:10.1007/s11205-011-9843.4.

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.

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*, 736–753. doi:10.1177/0013164498058005002.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37*, 113–115. doi:10.1037/0022-0167.37.1.113.

Mihura, J. L., Meyer, G. J., Dumitrascu, N., & Bombel, G. (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the comprehensive system. *Psychological Bulletin, 139*, 548–605. doi:10.1037/a0029406.

Qualls, A. L., & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. *Educational and Psychological Measurement, 56*, 209–214. doi:10.1177/0013164496056002002.

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable digit span: A systematic review and cross-validation study. *Assessment, 19*, 21–30. doi:10.1177/1073191111428764.

Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences & the Law, 31*, 55–73. doi:10.1002/bsl.2053.

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., & Maraun, M. D. (2009). Psychometric assessment and reporting practices: Incongruence between theory and practice. *Journal of Psychoeducational Assessment, 27*, 465–476.

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., Ferguson, L. P., Knudsen, J. R. S., & Legere, J. C. (2010). A review of psychometric assessment and reporting practices: An examination of measurement-oriented versus non-measurement-oriented domains. *Canadian Journal of School Psychology, 25*, 246–259. doi:10.1177/0829573510375549.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent Journal of Counseling & Development research articles. *Journal of Counseling & Development, 76*, 436–441.

Traub, R. E. (1994). *Reliability for the social sciences*. London: Sage.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *The Journal of Experimental Education, 67*, 335–341. doi:10.1080/00220979909598487.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509–522. doi:10.1177/00131640021970682.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62*, 562–569. doi:10.1177/0013164402062004002.

Van Remoortel, H., Giavedoni, S., Raste, Y., Burtin, C., Louvaris, Z., Gimeno-Santos, E., et al. (2012). Validity of activity monitors in health and chronic disease: A systematic review. *The International Journal of Behavioral Nutrition and Physical Activity, 9*, 84. doi:10.1186/1479-5868-9-84.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58*, 21–37. doi:10.1177/0013164498058001003.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. *Educational Researcher, 9*, 5–10. doi:10.2307/1175221.

Part IV
Health and Medicine

Chapter 12
Reporting of Measurement Validity in Articles Published in *Quality of Life Research*

Eric K.H. Chan, Bruno D. Zumbo, Michelle Y. Chen, Wen Zhang, Ira Darmawanti, and Olievia P. Mulyana

Health is not just the absence of disease, but "a state of complete physical, mental, and social well-being" (World Health Organization 1948). This definition suggests the importance of including the concept of well-being and quality of life (QoL) in assessing health. QoL in health is a broad, multidimensional concept that encompasses "general health, physical functioning, physical symptoms and toxicity, emotional functioning, cognitive functioning, role functioning, social well-being and functioning, sexual functioning and existential issues" (Fayers and Machin 2007, p. 4). Of late QoL has received increased attention in health research, practice, and policy.

Researchers, practitioners, and clinicians have increasingly recognized the importance of the assessment of QoL and it has become an important adjunct to traditional biomedical measures to assess or evaluate the burden of disease. For instance, individuals who suffer from chronic diseases have multifaceted healthcare-related needs and the diseases often affect the individuals' functioning and quality of life. These functioning and quality of life issues cannot be assessed via traditional biomedical methods. Psychometric instruments which enable these individuals to report their functioning and quality of life issues can allow researchers, clinicians, and policy makers to more effectively assess, monitor, and address the functioning and quality of life issues. Research has demonstrated that QoL information can improve collaboration between different disciplines in the health care system, improve health care plans, and improve communication between health care providers and patients (Detmar et al. 2002; Greenhalgh and

E.K.H. Chan (✉) • B.D. Zumbo, Ph.D. • M.Y. Chen • W. Zhang
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; bruno.zumbo@ubc.ca

I. Darmawanti • O.P. Mulyana
Department of Educational Psychology and Guidance, State University of Surabaya, Ketintang Baru XIV/2, Surabaya, East Java 60231, Indonesia

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_12,

Meadows 1999; Marshall et al. 2006; Velikova et al. 2004). It is therefore imperative to ensure the quality of psychometric QoL instruments.

In the area of health research there are two broad classes of measurement: psychometric and econometric. The former typically involves self-reports of health status, functioning, quality of life, symptoms, side effects, experience, and satisfaction, whereas the latter involves self-reports framed within an economic utility methodology (the utility approach to measurement is derived from decision-making theory and is common in health economics studies/evaluations.). Our focus is on the psychometric approach.

In the development and evaluation of a psychometric instrument, validity is a fundamental issue (AERA et al. 1999; Kane 2006; Messick 1989). The importance of validity is also reflected in policy documents, recommendations, and guidelines by a number of health organizations, such as the United States Food and Drug Administration (*Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*, published in 2009), the European Medicines Agency (*Reflection Paper on the Regulatory Guidance for the Use of Health-Related Quality of Life [HRQL] Measures in the Evaluation of Medicinal Products*, published in 2005), the Scientific Advisory Committee of the Medical Outcomes Trust (2002), and the joint Pharmaceutical Research and Manufacturers of America Health Outcomes Committee (PhRMA HOC) and the Division of Drug, Marketing, Advertising, and Communications of the FDA (DDMAC FDA) (Santanello et al. 2002). Task forces have also been established to develop guidelines on the best practices in the reporting of psychometric properties (including validity) of QoL instruments in clinical trials (Brundage et al. 2013; Calvert et al. 2013).

The theories of validity and methods for validation have become more advanced, expansive, and complex over the course of the last three decades. In addition to the traditional sources of validity such as content, relations to other variables (e.g., discriminant, and convergent validity), and internal structure, the contemporary view of validity contends that evidence based on response processes and consequences are emerging as important sources of validity evidence (AERA et al. 1999; Messick 1989; Hubley and Zumbo 2011). There is agreement among validity theorists that the integration and accumulation of validity evidence from various sources is needed to support the validity of the interpretation and inferences made from the scores arising from psychometric instruments (AERA et al. 1999; Kane 2006; Messick 1989; Zumbo 2007, 2009).

The purpose of the present study was to review the reporting of validity evidence in papers published in *Quality of Life Research*, with an aim towards investigating and improving the validation practices of health related QoL instruments. While the investigation of the reporting of validity evidence and validation practices have been conducted in education and psychology (e.g., Cizek et al. 2008, 2010), little is known in the area of QoL in health research. Examining the reporting characteristics is a useful strategy to investigate how a psychometric validation study in the area of QoL in health is designed.

Quality of Life Research is the official journal of the International Society for Quality of Life Research (ISOQOL). This scholarly society, and its official research

journal, was selected because, ISOQOL is internationally recognized as the authority for health-related quality of life research. With members from over 40 countries, ISOQOL is committed to high quality research on the measurement of health-related quality of life. Its mission is to advance the scientific study of health-related quality of life to enhance health care quality, identify effective health interventions, and improve people's health. We believe the papers published in that journal are authoritative resources for shaping research and serve as a fruitful ground from which to investigate psychometric validation practices in QoL in health research. Our focus was not to evaluate the quality of the psychometric instruments, but rather on informing validation practices. The following research question was addressed: What sources of validity evidences are reported in the validation of psychometric instruments published in the journal?

Methods

We conducted a systematic search using the official website of the journal in February 2013. We used the following keywords, "development" OR "measurement" OR "psychometric" OR "psychometrics" OR "valid" OR "validation" OR "validity", to search for articles. The search resulted in 2,416 articles in total. Retrieving and reviewing all validity and validation articles published in the journal since its inception would be ideal. However, due to the large number of articles and our limited resources, our team chose to only include and review papers published in the journal in 2012 and those that were in press at the time of our search (our search results included papers that were online first), with an explicit focus on papers with the term "valid", "validity", or "validation" in the title. This approach, in our opinion, allows us to see the very recent practice of validation in this journal and is a useful strategy to capture articles that are explicitly stated as validity and validation papers. A total of 34 articles were included and were coded. We chose to include only empirical validation studies and excluded opinion and editorial articles, reviews, systematic reviews, and meta-analyses, theoretical papers, and articles on guidelines, statistical data simulations, and methodological recommendations. We also excluded preference-based, utility, and related studies because these studies come from an econometric tradition of how one develops and "validates" instruments, and hence the language and framework are different from the psychometric approach (Kopec and Willison 2003; Richardson and Zumbo 2000). Conference abstracts were also excluded. The present review was therefore delimited to studies using the psychometric approach to validation.

We developed a coding sheet for the coding of the characteristics and validity evidence presented in each of the 34 selected papers published in this journal. Following the framework of validity stated in the most current version of the *Standards for Educational and Psychological Testing* available at the time of the conduct of this study (AERA et al. 1999) and building from previous research (e.g., Cizek et al. 2008, 2010), the following sources of validity evidence were coded:

face, content, construct, predictive, concurrent, convergent, discriminant, response processes, consequences, reliability, and other. The coding was based on what the authors of the articles reported and the procedures involved. For instance, if an author in the article explicitly reported "content validity", content validity was coded. If "think-aloud" process was presented in the paper to investigate people's responses yet the evidence was not explicitly stated as "response processes", it was still coded as response process evidence. In a similar vein, if factor analytic results were presented but were not explicitly called internal structure evidence in the paper, we still coded them as internal structure evidence. Four of the authors of the present study completed the coding of the validity evidence reported in the 34 papers published and in press in 2012 in this journal. To ascertain consistency in the coding, we randomly selected six articles (from the 34) and they were coding by all four individuals. The remaining 28 articles were equality divided therefore each of the four individuals coded seven of the 28 articles. In other words, each of the four individuals coded 13 (seven unique and six common) articles. In terms of the agreement among raters, disagreement occurred in 9 of the 60 multiple ratings (i.e., 6 articles for the 10 sources of validity that were coded); and within each case only one rater disagreed with the other three. There were no disagreements among the raters for the face, predictive, discriminant, consequences, and response process codes – in all others there was one rater who disagreed with the other three for only one or two of the articles in each coding category. Overall, we consider this a high consistency among our four raters. In terms the final coding for reporting validation practices, any disagreements in the coding results were reviewed and inconsistencies were resolved by the first author. During the coding stage, we found that one article was on the content validity methodology (Magasi et al. 2012) and the article was excluded. Therefore, the total number of articles included in the present study was 33.

Results and Conclusions

Reporting of Validity Evidence

The results of the present study showed that researchers conducting validation studies are not relying on only one sources of validity evidence in the exclusion of all other sources. As shown in Table 12.1, the number of sources of validity evidence reported per study ranged from zero to five, with a mode of two. Two studies reported zero sources of validity evidence (the authors did not refer to any source of validity evidence) yet it presented itself as a validation study. As presented in Table 12.2, internal structure, reported in over half of the papers, was the most frequently reported source of evidence to support the measurement structure and the consistency of the items of a QoL instrument. Slightly over half of the studies reported convergent validity or construct validity evidence. Of the

Table 12.1 Frequency of number of validity sources reported

Number of sources	Frequency	Percent
0	2	6.1
1	0	0
2	12	36.4
3	11	33.3
4	5	15.2
5	3	9.1
Total	**33**	**100**

Table 12.2 Sources of validity reported[a]

Source of validity	Number	Percent
Internal structure	20	60.6
Construct	19	57.6
Convergent	19	57.6
Discriminant	14	42.4
Concurrent	8	24.2
Content	4	12.1
Predictive	1	3.0
Response processes	0	0
Face	0	0
Consequences	0	0

[a]A paper can report more than one source of validity

19 studies that reported construct validity evidence, ten employed correlations to examine the association between instruments or variables that are of theoretical or clinical relevance, eight employed factor analysis, item response theory or Rasch modeling, two presented convergent and discriminant validity evidence (but presented the results as construct validity evidence), two employed multi-trait multi-method (MTMM) technique, and one conducted analysis of group differences. Discriminant validity, which serves as a baseline to compare convergent validity, was reported in about 40 % of the papers. About a quarter of the papers reported concurrent validity evidence. Four studies reported evidence on content validity. Examples of the methods employed in the content validation studies included cognitive interview, content analysis of patient responses, and face to face interviews. Predictive validity evidence was reported in only one (1 [3.0 %]) study. Response processes and consequences, which are emerging as an important source of validity evidence, were not reported in any of the 33 studies (see Table 12.2). Other reported validity sources included criterion, known-group, and internal/external validity.

In this study, we reviewed the validity evidence reported in papers published in *Quality of Life Research*, the official journal of ISOQOL. Our purpose is not to critique individual papers but rather use the papers published in the journal as a "data source" to document the prevalence of validation practices. The results revealed that the sources of validity evidence reported in the journal vary, and authors are not focusing on one source of validity evidence at the exclusion of all

others. Similar to the findings of *Value in Health* (chapter 15 of this book), another journal that publishes QoL research in health, evidence of internal consistency was the most widely reported source of validity evidence in *Quality of Life Research*. Other commonly reported sources of validity include those involving convergent and construct.

No evidence based on response processes or consequences are reported in our sample of papers, although these two sources are important in the process of validation (Hubley and Zumbo 2011, 2013; Messick 1989, 1995; Zumbo 2007, 2009). Neglecting these two important sources of validity evidence could affect the quality of QoL research in health which in turn may affect the quality of healthcare provided to people.

Response processes refer to the cognitive or thinking processes involved when an individual responds to items on a QoL instrument. The purpose is to investigate how and why people respond to items on QoL instruments the way they do. Although the importance of examining this substantive aspect of validity has been stressed (AERA et al. 1999; Messick 1989, 1995) and is emerging as central to validity claims (Hubley and Zumbo 2011, 2013; Messick 1995; Zumbo 2007, 2009), we found that this source of validity evidence was not present in any of the 33 papers published in *Quality of Life Research*.

Response processes also concerns the influence individual characteristics have on their response to QoL items. For example, when asked about emotional functioning, it is important that people are reasoning about their emotional functioning and their responses should not be influenced by characteristics such as social desirable responding or stigma associated with poor emotional functioning. These issues may lead to invalid QoL assessment results.

Although the examination of consequences is emerging as central to validity claims (Hubley and Zumbo 2011, 2013; Messick 1995; Zumbo 2007, 2009), consequences were not reported in any of the 33 papers we reviewed in this study. Consequences include both positive and negative consequences of the intended use of test scores (AERA et al. 1999; Hubley and Zumbo 2011, 2013; Messick 1989). First, intended use is the claims or decisions we want to make based on the scores on a QoL instrument and is part of the entire validation activities. Issues such as construct underrepresentation and irrelevant variance can negatively influence the intended use of a QoL instrument. For instance, research in mental health has shown that males tend to have lower scores (i.e., better mental health) than females. However, differential item functioning (DIF) research has demonstrated gender DIF, suggesting that males are less likely to endorse certain mental health items when the level of mental health between the two genders are controlled for. This lack of invariance may shed light on why males tend to appear to have better mental health and may weaken the intended interpretation of the difference in mental health scores between males and females. Such findings may also affect the validity of QoL assessment in health.

The misuse of test scores is also a central concern although not a source of invalidity, per se (Hubley and Zumbo 2011, 2013). That is, misuse, in and of itself, does not invalidate the appropriate use of an instrument. An example of misuse is in the diagnosis of clinical depression. A diagnosis of clinical depression cannot be

made based on the scores arising from screening instruments; additional clinical evaluation is needed (Maurer 2012; Pignone et al. 2002; Sharp and Lipsky 2002). As the intended use of screening instruments is to identify individuals who *may* have clinical depression and to identify those who *may* require additional mental health evaluation, but not to make official diagnosis. Making a diagnosis of depression based solely on the scores on a depression screening instrument is an example of misuse. Such a misuse may result in over- or under-diagnosis of clinical depression. This may have negative consequences on epidemiological findings, diagnostic decisions, and even insurance coverage. A misuse of the depression screening instrument as a clinical diagnostic tool does not invalidate it as a screen instrument – rather it just invalidates it as a diagnostic tool.

The issue of consequences in validity is mentioned in task force reports by some health care associations. For instance, this issue is included in the Ad Hoc Task Force Report on the incorporation of patient perspective into drug development and communication (Acquadro et al. 2003) by the International Society for Pharmacoeconomics and Outcomes Research Patient Reported Outcomes Harmonization Group. Although the term consequence was not stated in the report explicitly, the report stated that the inclusion of QoL assessment in clinical studies should be made with the intended claims in mind. This suggests that the issue of consequences need to be considered when using psychometric QoL instruments in clinical research. Our findings that consequences were not reported in any of the 33 papers we reviewed in this study suggests that more effort is needed to promote the inclusion of consequences of appropriate use of an instrument in validation activities in QoL in health, and not just the misuse of a measure. The formation of task forces and development of best practices and reporting guidelines may be helpful in promoting the inclusion of consequences in validation practices in QoL research in health.

It is important to note that we are not suggesting that researchers conducting validation work are required to have all five sources of validity evidence in all cases. Instead, the use of the instrument and the patient population should drive the sources of validity evidence needed to support the score interpretation. Our review of the 33 articles published in *Quality of Life Research* shows that it is not common for validation researchers to use the *Test Standards*, or theoretical frameworks by Messick or Kane to guide their validation practices. At present, the lack or theoretical framework to guide validation practices makes validation activities seem like "stamp collecting" in which a few sources (sometimes only a single source) of validity are collected to support the validity of score interpretation. This is in contrast with the currently accepted view in validity theory that an integrated evidential basis is needed to support validity claims.

It is also interesting to note that only a very small percentage (3.0 %) of the articles we reviewed in this study reported predictive validity evidence. It seems reasonable for future research in QoL to place more emphasis on the investigation of predictive validity. Such effort has the potential to increase the value of including QoL assessment in health research and care. Our findings also suggest that researchers investigating the predictive values of QoL in health may not be using the psychometric language (e.g., predictive validity). In cancer research, for

instance, QoL predicts survival (e.g., Efficace et al. 2006; Karvonen-Gutierrez et al. 2008; Maione et al. 2005; Montazeri 2009), but the term predictive validity is not commonly used. It seems reasonable for future research in QoL to place more emphasis on the investigation of predictive validity. Such effort has the potential to increase the value of including QoL assessment in health research and care. Researchers conducting research syntheses should be mindful that the search terms used to identify relevant studies could impact the quality of the synthesis. For example, if the search term "predictive validity" is used, the results may exclude studies in which predictive validity was indeed examined, yet the authors examined it using a different term.

The present study has limitations. First, we focused on examining the 33 articles published in *Quality of Life Research* in 2012. Given the relatively limited focus, the results may not be generalizable to all of the validation studies published in the journal. As mentioned, although it would be ideal to review all validity and validation articles published in the journal since the inception of the journal, due to the large number of articles and our limited resources, our focus on articles published very recently would still give us a good understanding of the validity evidence reported and validation practices in the area of QoL in health. Second, just because the authors did not report all sources of validity evidence does not mean no such evidence have been done or that authors are not aware of the other sources of validity, such as consequences and response processes. The design of the present study did not allow us to examine the reasons behind the reporting of the validity evidence in papers published in *Quality of Life Research*. Future research is needed to investigate this issue.

References[1]

Acquadro, C., Berzon, R., Dubois, D., Leidy, N. K., Marquis, P., Revicki, D., et al. (2003). Incorporating patient's perspective into drug development and communication: An ad hoc task force report of the patient-reported outcomes (PRO) harmonization group meeting at the Food and Drug Administration, February 16, 2001. *Value in Health, 6*, 522–531.

*Al-Janabi, H., Peters, T. J., Brazier, J., Bryan, S., Flynn, T. N., Clemens, S., Moody, A., & Coast, J. (2013). An investigation of the construct validity of the ICECAP-A capability measure. *Quality of Life Research, 22*, 1831–1840.

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.

*Ashing-Giwa, K., & Rosales, M. (2013). A cross-cultural validation of patient-reported outcomes measures: A study of breast cancers survivors. *Quality of Life Research, 22*, 295–308.

*Baroin, A., Chopard, G., Siliman, G., Michoudet, C., Vivot, A., Vidal, C., Mokadym, H., Lavier, A., Berger, E., Rumbach, L., & Rude, N. (2013). Validation of a new quality of life scale related to multiple sclerosis and relapses. *Quality of Life Research, 22*, 1943–1954.

[1] References marked with an asterisk indicate studies included in the review.

*Bartram, D. J., Sinclair, J. M., & Baldwin, D. S. (2013). Further validation of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) in the UK veterinary profession: Rasch analysis. *Quality of Life Research, 22*, 379–391.

Brundage, M., Blazeby, J., Revicki, D., Bass, B., de Vet, H., Duffy, H., et al. (2013). Patient-reported outcomes in randomized clinical trials: Development of ISOQOL reporting standards. *Quality of Life Research, 22*, 1161–1175.

Calvert, M., Blazeby, J., Altma, D. G., Revicki, D. A., Moher, D., & Brundage, M. D., for the CONSORT PRO Group. (2013). Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. *Journal of the American Medical Association, 309*, 814–822.

*Cho, S., Kim, H. Y., & Lee, J. H. (2013). Validation of the Korean version of the Pain Catastrophizing Scale in patients with chronic non-cancer pain. *Quality of Life Research, 22*, 1767–1772.

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.

Detmar, S. B., Muller, M. J., Schornagel, J. H., Wever, L. D. V., & Aaronson, N. K. (2002). Health-related quality-of-life assessments and patient-physician communication: A randomized controlled trial. *Journal of the American Medical Association, 228*, 3027–3034.

Efficace, F., Bottomley, A., Coens, C., van Steen, K., Conroy, T., Schöffski, P., et al. (2006). Does a patient's self-reported health-related quality of life predict survival beyond key biomedical data in advanced colorectal cancer? *European Journal of Cancer, 42*, 42–49.

European Medicines Agency. (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: European Medicines Agency.

Fayers, P., & Machin, D. (2007). *Quality of life: The assessment, analysis and interpretation of patient-reported outcomes* (2nd ed.). Chichester: Wiley.

*Ferreira, N. B., Eugenicos, M. P., Morris, P. G., & Gillanders, D. T. (2013). Measuring acceptance in irritable bowel syndrome: Preliminary validation of an adapted scale and construct utility. *Quality of Life Research, 22*, 1761–1766.

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.

*Frans, F. A., van Wijngaarden, S. E., Met, R., & Koelemay, M. J. W. (2012). Validation of the Dutch version of the VascuQol questionnaire and the Amsterdam linear disability score in patients with intermittent claudication. *Quality of Life Research, 21*, 1487–1493.

*Franz, M., Fritz, M., & Meyer, T. (2013). Discriminant and convergent validity of a subjective quality-of-life instrument aimed at high content validity for schizophrenic persons. *Quality of Life Research, 22*, 1113–1122.

*Gonçalves, R. S., Gil, J. N., Cavalheiro, L. M., Costa, R. D., & Ferreira, P. L. (2012). Reliability and validity of the Portuguese version of the Stroke Impact Scale 2.0 (SIS 2.0). *Quality of Life Research, 21*, 691–696.

Greenhalgh, J., & Meadows, K. (1999). The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: A literature review. *Journal of Evaluation in Clinical Practice, 5*, 401–416.

*Holzhausen, M., & Martus, P. (2013). Validation of a new patient-generated questionnaire for quality of life in an urban sample of elder residents. *Quality of Life Research, 22*, 131–135.

*Horsman, S., Olson, K., Au, H., & Ghosh, S. (2012). Symptom assessment in ambulatory oncology: Initial validation of the nurse-developed Modified Ambulatory Care Flow Sheet (MACFS). *Quality of Life Research, 21*, 899–908.

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.

*Huijier, H. A., Sagherian, K., & Tamim, H. (2013). Validation of the Arabic version of the EORTC quality of life questionnaire among cancer patients in Lebanon. *Quality of Life Research, 22*, 1473–1481.

*Hunger, M., Sabariego, C., Stollenwerk, B., Cieza, A., & Leidi, R. (2012). Validity, reliability and responsiveness of the EQ-5D in German stroke patients undergoing rehabilitation. *Quality of Life Research, 21*, 1205–1216.

*Jankovic, S., Vukicevic, J., Djordjevic, S., Jankovic, J., Marinkovic, J., & Basra, M. K. (2013). The Cardiff Acne Disability Index (CADI): Linguistic and cultural validation in Serbian. *Quality of Life Research, 22*, 161–166.

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

Karvonen-Gutierrez, C. A., Ronis, D. L., Fowler, K. E., Terrell, J. E., Gruber, S. B., & Duffy, S. A. (2008). Quality of life scores predict survival among patients with head and neck cancer. *Journal of Clinical Oncology, 26*, 2754–2760.

*Kleinman, L., Benjamin, K., Viswanathan, H., Mattera, M. S., Bosserman, L., Blayney, D. W., & Revicki, D. A. (2012). The anemia impact measure (AIM): Development and content validation of a patient-reported outcome measure of anemia symptoms and symptom impacts in cancer patients receiving chemotherapy. *Quality of Life Research, 21*, 1255–1266.

*Knibb, R. C., & Stalker, C. (2012). Validation of the food allergy quality of life-parental burden questionnaire in the UK. *Quality of Life Research, 22*, 1841–1849.

Kopec, J. A., & Willison, K. D. (2003). A comparative review of four preference-weighted measures of health-related quality of life. *Journal of Clinical Epidemiology, 56*, 317–325.

*Krägeloh, C. U., Kersten, P., Rex Billington, D., Hsu, P. H., Shepherd, D., Landon, J., & Feng, X. J. (2013). Validation of the WHOQOL-BREF quality of life questionnaire for general use in New Zealand: Confirmatory factor analysis and Rasch analysis. *Quality of Life Research, 22*, 1451–1457.

*Lafaye, A., De Chalvron, S., Houédé, N., Eghbali, H., & Cousson-Gélie, F. (2013). The Caregivers Quality of Life Cancer index scale (CQoLC): An exploratory factor analysis for validation in French cancer patients' spouses. *Quality of Life Research, 22*, 119–122.

*Landgraf, J. M., Vogel, I., Oostenbrink, R., van Baar, M. E., & Raat, H. (2013). Parent-reported health outcomes in infants/toddlers: Measurement properties and clinical validity of the ITQOL-SF47. *Quality of Life Research, 22*, 635–646.

*Launois, R., Le Moine, J. G., Lozano, F. S., & Mansilha, A. (2012). Construction and international validation of CIVIQ-14 (a short form of CIVIQ-20), a new questionnaire with a stable factorial structure. *Quality of Life Research, 21*, 1051–1058.

*Löve, J., Moore, C. D., & Hensing, G. (2012). Validation of the Swedish translation of the general self-efficacy scale. *Quality of Life Research, 21*, 1249–1253.

*Lucas-Carrasco, R. (2012). The WHO quality of life (WHOQOL) questionnaire: Spanish development and validation studies. *Quality of Life Research, 21*, 161–165.

Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R. D., Brod, M., Snyder, C., Boers, M., & Cella, D. (2012). Content validity of patient-reported outcome measures: Perspectives from a PROMIS meeting. *Quality of Life Research, 21*, 739–746.

Maione, P., Perrone, F., Gallo, C., Manzione, L., Piantedosi, F., Barbera, S., et al. (2005). Pretreatment quality of life and functional status assessment significantly predict survival of elderly patients with advanced non-small-cell lung cancer receiving chemotherapy: A prognostic analysis of the Multicenter Italian Lung Cancer in the Elderly Study. *Journal of Clinical Oncology, 23*, 6865–6872.

Marshall, S., Haywood, K., & Fitzpatrick, R. (2006). Impact of patient-reported outcome measures on routine practice: A structured review. *Journal of Evaluation in Clinical Practice, 12*, 559–568.

Maurer, D. M. (2012). Screening for depression. *American Family Physician, 85*, 139–144.

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

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

Montazeri, A. (2009). Quality of life data as prognostic indicators of survival in cancer patients: An overview of the literature from 1982 to 2008. *Health and Quality of Life Outcomes, 7*, 102.

*Monticone, M., Baiardi, P., Ferrari, S., Foti, C., Mugnai, R., Pillastrini, P., Rocca, B., & Vanti, C. (2012). Development of the Italian version of the Pain Catastrophising Scale (PCS-I): Cross-cultural adaptation, factor analysis, reliability, validity and sensitivity to change. *Quality of Life Research, 21*, 1045–1050.

*Monticone, M., Ferrante, S., Giorgi, I., Galandra, C., Rocca, B., & Foti, C. (2013). Development of the Italian version of the 42-item Chronic Pain Coping Inventory, CPCI-I: Cross-cultural adaptation, factor analysis, reliability and validity. *Quality of Life Research, 22*, 1459–1465.

*Ostini, R., Dower, J., & Donald, M. (2012). The Audit of Diabetes-Dependent Quality of Life 19 (ADDQoL): Feasibility, reliability and validity in a population-based sample of Australian adults. *Quality of Life Research, 21*, 1471–1477.

Pignone, M. P., Gaynes, B. N., Rushton, J. L., Burchell, C. M., Orleans, C. T., Mulrow, C. D., & Lohr, K. N. (2002). Screening for depression in adults: A summary of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine, 136*, 765–776.

Richardson, C. G., & Zumbo, B. D. (2000). A statistical examination of the Health Utility Index-Mark III as a summary measure of health. *Social Indicators Research, 51*, 171–191.

Santanello, N. C., Baker, D., & Cappelleri, J. C. (2002). Regulatory issues for health-related quality of life – PhRMA Health Outcomes Committee Workshop, 1999. *Value in Health, 5*, 14–25.

*Sarkin, A. J., Groessl, E. L., Carlson, J. A., Tally, S. R., Kaplan, R. M., Sieber, W. J., & Ganiats, T. G. (2013). Development and validation of a mental health subscale from the Quality of Well-Being Self-Administered. *Quality of Life Research, 22*, 1685–1696.

*Sawatzky, R., Ratner, P. A., Kopec, J. A., & Zumbo, B. D. (2012). Latent variable mixture models: A promising approach for the validation of patient reported outcomes. *Quality of Life Research, 21*, 637–650.

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.

Sharp, L. K., & Lipsky, M. S. (2002). Screening for depression across the lifespan: A review of measures for use in primary care settings. *American Family Physician, 66*, 1001–1008.

*Stevanovic, D., Tadic, I., Novakovic, T., Kisic-Tepavcevic, D., & Ravens-Sieberer, U. (2013). Evaluating the Serbian version of the KIDSCREEN quality-of-life questionnaires: Reliability, validity, and agreement between children's and parents' ratings. *Quality of Life Research, 22*, 1729–1737.

*ten Klooster, P. M., Taal, E., Oostveen, J. C., Harmsen, E. J., Tugwell, P. S., Rader, T., Lyddiatt, A., & van de Laar, M. A. (2013). Translation and validation of the Dutch version of the Effective Consumer Scale (EC-17). *Quality of Life Research, 22*, 423–429.

*Uysal, M. A., Mungan, D., Yorgancioglu, A., Yildiz, F., Akgun, M., Gemicioglu, B., Turktas, H., & The Turkish Asthma Control Test (TACT) Study Group. (2013). The validation of the Turkish version of Asthma Control Test. *Quality of Life Research, 22*, 1773–1779.

Velikova, G., Booth, L., Smith, A. B., Brown, P. M., Lynch, P., Brown, J. M., & Selby, P. J. (2004). Measuring quality of life in routine oncology practice improves communication and patient well-being: A randomized controlled trial. *Journal of Clinical Oncology, 22*, 714–724.

World Health Organization. (1948). Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19–22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.

*Yoo, H. J., Kim, S. B., Yoon, D. H., Park, S. I., Kim, J. H., Cella, D., Jung, H. Y., Lee, G. H., Choi, K. D., Song, H. J., Song, H. Y., Shin, J. H., & Cho, K. J. (2012). Translation and validation of Korean Functional Assessment of Cancer Therapy-Esophageal (FACT-E) scale with squamous cell carcinoma and chemoradiation-only patients. *Quality of Life Research, 21*, 1451–1457.

*Yuksel, H., Yilmaz, O., Dogru, D., Karadag, B., Unal, F., & Quittner, A. L. (2013). Reliability and validity of the Cystic Fibrosis Questionnaire-Revised for children and parents in Turkey: Cross-sectional study. *Quality of Life Research, 22*, 409–414.

*Zhao, H. P., Liu, Y., Li, H. L., Ma, L., Zhang, Y. J., & Wang, J. (2013). Activity limitation and participation restrictions of breast cancer patients receiving chemotherapy: Psychometric properties and validation of the Chinese version of the WHODAS 2.0. *Quality of Life Research, 22*, 897–906.

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/Boston: 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: Information Age Publishing.

Chapter 13
Validity Evidence for a Perceived Social Support Measure in a Population Health Context

Daniel W. Cox and Jess J. Owen

Measurement validity refers to the extent to which the meaning or interpretation that is attributed to a score is accurate (Messick 1995). Further, validity evidence is tied to the specific purpose(s) of score interpretation (American Educational Research Association American Psychological Association and National Council on Measurement in Education 1999). For this reason, when scores are used for purposes other than what they were originally intended, validity evidence for this new purpose is necessary to empirically and theoretically support score interpretations. If validity evidence is absent, inaccurate interpretations may be made.

It is best to think of validation as a process rather than a goal; it is iterative (Messick 1995). Validity should be evaluated for every new purpose/interpretation of a score. Notably, this includes "across persons or population groups and across settings or contexts" (Messick 1995, p. 741). Thus, it is important that the generalizability of samples be considered when evaluating validity evidence. For example, when describing validity evidence of a score on a perceived social support measure, to write that 'concurrent criterion validity was demonstrated in a sample of adult men and women,' has different implications than writing that 'concurrent criterion validity was demonstrated in a sample of 18- to 23-year-old Israeli men and women actively serving in the Israeli Army.' It is likely that the age, ethnicity, and occupation/context of the sample influence the score's meaning. Thus, the score on a measure may have different meaning in a sample of 70- to 79-year-old widowers recruited from a physical rehabilitation center in Beijing compared with a sample of gifted second-grade boys and girls residing in rural Pennsylvania. Contrarily, it may have the same meaning across these groups; by assessing validity between populations and contexts, score generalizability is being empirically evaluated.

D.W. Cox (✉) • J.J. Owen
Counseling Psychology 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: dan.cox@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_13,

Population Health Research

Population health researchers study the role of health determinants (i.e., risk and protective factors) on health outcomes (Kindig and Stoddart 2003). There is substantial latitude when defining health determinants, health outcomes, and populations, which depends on the research questions being asked. Further, they study the relation of these variables, both within and between populations. Notably, population health research differs from traditional individual-differences psychological or educational research insofar as the focus is on understanding populations rather than persons.[1] Thus, we would expect validation practices to reflect this difference – the populations' score validity is of interest rather than the persons'.

To evaluate populations, population health researchers apply sophisticated sampling methods so their samples represent populations of interest. Thus, their conclusions can apply to entire populations without assessing every person within a population. For example, Statistics Canada's Canadian Community Health Survey (CCHS) is an ongoing survey of health-related practices in Canada (Statistics Canada 2011). To attain a representative sample of the country's residents; provinces, territories, and health regions (i.e., provincial subdivisions) are sampled based on the number of people residing in those areas. Using this method, a *representative sample* is attained – the sample of 65,000 participants is representative of the 35,000,000 people in the national population. To further facilitate accurate representativeness, each participant is given a weight (i.e., the number of people within the total population that person represents). Using sampling weights allows researchers to ask population-level research questions without sampling the entire population. Thus, research questions can be asked of the entire population, 'Among all Canadians, does perceived social support influence their likelihood of seeking professional mental health treatment?,' or subpopulations, 'Among Aboriginal adolescent males living in British Columbia, does perceived social support influence their likelihood of seeking professional mental health treatment?.' While there is variation in population health sampling methods, this example illustrates the context in which measurement validity can be evaluated by population health researchers.

The Medical Outcomes Study-Social Support Survey (MOS-SS)

The MOS-SS is a 19-item self-report measure of perceived social support (Sherbourne and Stewart 1991). It was developed to measure different functions (i.e., functional aspects) of perceived social support. These functions were based on

[1] The authors note that the boundaries of these fields are not concrete as findings from psychological and educational research may apply to populations.

reviewing the literature and identifying functions regarded as most important in empirically-based models of social support. This resulted in the generation of 50 items, and following a variety of empirical item-reduction techniques, was reduced to 19 items representing four functions of perceived social support: (a) emotional/informational (e.g., having people to talk to about problems and get advice from), (b) tangible (e.g., having people who give material aid), (c) affectionate (e.g., having people who give love and affection), and (d) positive social interaction (e.g., having people to spend time with). Respondents indicate how often each of the support functions is available to them from 1 (*none of the time*) to 5 (*all of the time*). Items within each function can be summed for a subscale score and all of the items can be summed for an overall index of perceived social support.

The MOS-SS was developed in the context of evaluating both process and outcome variables in a longitudinal study of patients with chronic medical conditions (Sherbourne and Stewart 1991). Specifically, the study was designed to examine how physicians' practices and different health care systems influenced patients' outcomes. Thus, original validity evidence was examined with 2,987 patients who were part of the Medical Outcome Study. Since, the measure has been adopted by population health researchers and used in several large-scale population health surveys (Statistics Canada 2011).

Study Purpose

The purpose of the present study was to evaluate the validity evidence for the MOS-SS. Specifically, we were interested in if validity evidence reflected the use of the MOS-SS in population health research rather than individual differences research. We had three major research questions: (a) What were the characteristics of the samples in which validity was evaluated? (b) Were samples representative of populations? and (c) What sources of validity evidence have been evaluated?

Methods

Sampling

We conducted a systematic literature review of peer-reviewed journal articles that evaluated the validity of MOS-SS scores. We used PsycINFO, MEDLINE, and ERIC and searched for keywords in the title and abstract. Search words were "MOS social support," "Medical Outcomes Study Social Support," and "valid*". This resulted in 69 peer-reviewed articles that were then screened to determine if they evaluated the validity of the MOS-SS. From the 69 articles, we concluded that

20 explicitly evaluated the validity of the MOS-SS. From these 20, 4 were removed – 2 because they were not in English and published in journals that we could not attain and 2 because the authors used the MOS-SS to examine the validity of other measures and did not examine the validity of the MOS-SS – resulting in 16 peer-reviewed journal articles. One article consisted of three samples and validity was examined separately in each (Moser et al. 2012), so we coded it as three separate studies. Thus, we coded a total of 18 studies.

Coding

All articles were reviewed by the two authors and coded via a structured coding form (see Table 13.1). The first author had doctoral-level course work in the area of psychometrics and had published in the area; the second author was a graduate student with undergraduate and graduate level course work in psychometrics, psychological testing, research methods, and statistics. Determinations about the types of validity evaluated were made by the current authors in-line with *The Standards* (AERA and APA 1999). Thus, if a study's authors stated that they evaluated one type of validity, yet in the description of their methods the type of validity they evaluated — according to *The Standards* — was a different type, we coded consistent with *The Standards*. Because most articles presented more than one type of validity evidence, those that did were counted in each appropriate category (i.e., more than once). Particular emphasis was given to identifying the populations that samples were drawn from and if those samples were representative of those populations – issues particularly important in population health research. Coding was done independently by each author, who met and discussed discrepancies so that 100 % agreement on all variables was reached.

Results

Overview of Studies Coded

Of the 18 studies coded, the majority evaluated the 19-item MOS-SS (66.7 %), followed by the 8-item Modified MOS-SS (mMOS-SS) (16.7 %), and the 12-item (5.6 %), 4-item (5.6 %), and Computer Mediated MOS-SS (5.6 %). Most of the investigations were psychometric (94.4 %) with one (5.6 %) being in the context of a research study. Further, none (0 %) of the authors indicated that their validity perspective came from *The Standards* or from Messick's work.

Table 13.1 Coding form

Information coded	Coding options (if applicable)
Age	
Sex	
Ethnicity	
Sample size	
Form	
Language of form	
Type of investigation	(a) Psychometric
	(b) Research
Validity perspective/theory used by the study's authors	(a) *The Standards*
	(b) Messick
	(c) Unsure/not clear
Representative of population	(a) Yes
	(b) No
Authors stated they evaluated content validity	(a) Yes
	(b) No
Factor analysis	(a) Yes
	(b) No
Item interrelationships	(a) Yes
	(b) No
Invariance	(a) Yes
	(b) No
Mokken scaling	(a) Yes
	(b) No
Simplex pattern	(a) Yes
	(b) No
Convergent	(a) Yes
	(b) No
Criterion predictive	(a) Yes
	(b) No
Criterion concurrent	(a) Yes
	(b) No
Generalizations	(a) Yes
	(b) No
Discriminant	(a) Yes
	(b) No
Known group	(a) Yes
	(b) No
Nomological networks	(a) Yes
	(b) No
Construct	(a) Yes
	(b) No
Response processes	(a) Analysis of individual responses by interview
	(b) Examining similarities/differences in responses by distinct groups or investigations in which researchers collect, record, and interpret data
	(c) No

(continued)

Table 13.1 (continued)

Information coded	Coding options (if applicable)
Consequences	(a) Benefits associated with test use
	(b) Negative uses
	(c) Other
	(d) Unsure/not clear
	(e) No

Research Question 1: What Were the Characteristics of the Samples in Which Validity Was Evaluated?

A full breakdown of the populations sampled is presented in Table 13.2. Of the 18 studies reviewed, the majority evaluated adult samples (66.7 %), followed by older adults (27.8 %), and young adults (5.6 %). Further, validity was evaluated in either mixed sex (66.7 %) or female (33.3 %) samples, no studies (0 %) evaluated MOS-SS validity in solely male samples. Multiple validity studies were conducted in the U.S. (38.9 %) and China/Taiwan (22.2 %), while several countries had one validity study (i.e., Portugal, Italy, Hong Kong, Malaysia, Canada, Brazil, and South Africa). Further, validity studies with the MOS-SS translated into several languages were conducted. Languages included English (50 %), Chinese (dialect not specified) (27.8 %), Portuguese (5.6 %), Brazilian Portuguese (5.6 %), French (5.6 %), Italian (5.6 %), and Malay (5.6 %). Regarding race/ethnicity, few studies evaluated validity evidence in specific races/ethnicities (11.1 %); most (88.9 %) omitted race/ethnicity from their inclusion/exclusion criteria. The majority of the studies were conducted with clinical populations (77.8 %), most of them included heart disease patients (27.8 %) or breast cancer survivors (16.7 %), while the minority (22.2 %) were conducted among nonclinical populations.

Research Question 2: Were Samples Representative of Populations?

Of the studies coded, only one (5.6 %) evaluated the validity of MOS-SS scores in a representative sample (Robitaille et al. 2011) (see Table 13.2). That study was conducted among a representative sample of older adult Canadians, using the English and French translations of the measure.

Table 13.2 Populations examined

Age	Sex	Race/ethnicity[a]	Representative of population	Clinical	Country	Language
Adults	Female	African or Latina-American	No	Breast cancer survivors	U.S.	English
Adults	Mixed	Non-specific	No	Chronic illness patients	Portugal	Portuguese
Young adults	Mixed	Non-specific	No	Non-clinical	Italy	Italian
Adults	Female	Non-specific	No	Mothers with child in mental health treatment	U.S.	English
Older adults	Mixed	Non-specific	No	Heart disease patients	Hong Kong	Chinese
Adults	Female	Non-specific	No	4–12 weeks postpartum	Malaysia	Malay
Older adults	Female	Non-specific	No	Breast cancer survivors	U.S.	English
Adults	Female	Non-specific	No	Breast cancer survivors	U.S.	English
Adults	Female	Non-specific	No	Non-clinical	U.S.	English
Older adults[b]	Mixed	Non-specific	No	Non-clinical	U.S.	English
Older adults	Mixed	Non-specific	Yes	Non-clinical	Canada	English & French
Adults	Mixed	Non-specific	No	Hypertension, heart disease, diabetes, or depression patients	U.S.	English
Adults	Mixed	Non-specific	No	Family caregivers of cancer patients	Taiwan	Chinese
Adults[c]	Mixed	Non-specific	No	Hodgkin's lymphoma survivors	Brazil	Brazilian Portuguese
Adults	Mixed	Non-specific	No	Heart disease patients	China	Chinese
Adults	Mixed	Non-specific	No	Heart disease patients	China	Chinese
Adults	Mixed	Black South Africans	No	Diabetes patients	South Africa	English
Older adults	Mixed	Non-specific	No	Heart disease patients	China	Chinese

Note. Unless otherwise stated, adults = 18+, young adults = undergraduate college students, Older adults = 55 +

[a]Noted as non-specific if the authors omitted specific races/ethnicities in their inclusion/exclusion criteria

[b]Participants resided in a retirement community and were members of the computer club

[c]Defined adults as 16 +

Table 13.3 Frequencies of validity evidence sources

Source of evidence	Number of studies
Content	6
Internal structure	
Factor analysis	17
Item interrelationships	2
Invariance	1
Mokken scaling	1
Other: Simplex pattern	0
Relations to other variables	
Convergent	2
Criterion-predictive	0
Criterion-concurrent	15
Criterion-group differences	3
Generalizations	0
Discriminant	5
Nomological network	0
Construct validity	14
Response processes	0
Consequences	0

Note. Several studies reported more than one category of validity evidence; therefore, they were counted in each appropriate category (i.e., more than once)

Research Question 3: What Sources of Validity Evidence Have Been Evaluated?

For a full description of the types of validity evidence presented in the reviewed articles, see Table 13.3. For internal structures, the types of validity evidence reported included factor analysis (94.4 %), item interrelationships (11.1 %), invariance (5.6 %), and Mokken scaling (5.6 %). Regarding relations to other variables, the most common was criterion-concurrent (83.3 %), discriminant (27.8 %), criterion-group differences (16.7 %), and convergent (11.1 %). Further, 33.3 % reported content validity. No authors (0 %) reported criterion-predictive validity, response processes, or consequences.

Discussion

The purpose of the present study was to evaluate validation practices and validity evidence for the MOS-SS. To do this, we systematically reviewed the validity research on the measure. The MOS-SS was originally designed for research on patient reported outcomes and has since been adopted by population health researchers. Below, we discuss the validation practices reviewed and their strengths

and limitations in the context of population health research. Further, we offer suggestions for future directions in population health validation practices.

Evaluating Validity Within and Between Populations

One of the strengths of the validity evidence for the MOS-SS is that validity was evaluated among several international populations and in several different languages. Further, a variety of clinical samples and several non-clinical samples were used in the validity studies as well as mixed gender and solely female samples. While validity evidence for the MOS-SS was examined in several different populations, only one study directly evaluated the measure's factor structure between populations (Robitaille et al. 2011). In their study, the authors used measurement invariance to examine if the MOS-SS had a uniform factor structure for English and French speaking Canadians who completed the measure in their primary language. By using measurement invariance (i.e., measurement equivalence), researchers assess measures' validity between populations of interest by examining if the internal structure of the measure is the same in different populations (Vandenberg and Lance 2000). If it is not the same, the measure is considered biased (Brown 2006).

For many researchers, evaluating measurement invariance is difficult because it requires large samples from each population of interest to make between population conclusions. In the context of population health research, large sample sizes are common. Thus, for many population health researchers, samples are well suited for evaluating measurement invariance and are a practical way to evaluate validity between populations of interest. While measurement invariance is used to evaluate the consistency of a measure's internal structure between populations, structural consistency does not imply that each scale is measuring the same construct in each population. Thus, other evaluations of validity are necessary to determine if validity is equivalent between populations. That being said, we believe that evaluating measurement invariance is an important step for population health researchers examining measurement validity across populations of interest.

Multilevel Validity Evidence

Methodologists have argued that group-based (e.g., aggregate- or population-level) inferences require group-based validity (Zumbo and Forer 2011). Most of the history of the social and health sciences is made-up of individual-based inferences; thus, validity is almost exclusively conceptualized within an individual-differences paradigm. However, interest in population-level research questions and advances in research methods have facilitated multilevel validity evaluation. A common example of a multilevel context comes from education – students (level one) are nested within classrooms (level two), classrooms are nested within schools (level three),

and schools are nested within districts (level four). An inference at the student level – wealthier students score better on standardized achievement tests – has different implications than a similar inference at the district level – wealthier districts score better on standardized achievement tests. Further, we cannot assume that measures of socioeconomic status or academic achievement are equivalently valid at the individual and the district level. This framework is analogous in population health – for example, persons (level one) are nested within neighborhoods (level two), neighborhoods are nested within health regions (level three), health regions are nested within provinces (level four), and provinces are nested within countries (level five). Similar to the education example above, an inference at the person level – people with greater social support are less likely to have major depressive disorder – has different implications than a similar inference at the provincial level – provinces with greater social support have lower rates of major depressive disorder. Also, assuming measures of social support and major depressive disorder are equally valid at the person and provincial level is an assumption that requires empirical examination. None of the MOS-SS validation studies applied multilevel analysis. While it is beyond the scope of this chapter to fully explicate the theory and proposed methods of multilevel validity analysis (see Forer and Zumbo 2011; Zumbo and Forer 2011), we hope this brief discussion of the problem encourages population health researchers to consider multilevel conceptualizations and analyses, which we believe will improve the validity of population health measures, results, and inferences.

Representativeness of Samples

Validity evidence for the MOS-SS has been evaluated in nine countries on five continents. Presently, only one study examined the MOS-SS in a sample representative of a population – English and French speaking, older adult, Canadians (Robitaille et al. 2011). While validity studies were conducted in several countries, none used representative samples of the country's population. Nor did they examine representative samples of subpopulations within those countries. Therefore, while findings from these studies are encouraging, regarding MOS-SS score validity across populations, they fall short of assessing validity at a population level.

Relations to Other Variables

We found that more studies (i.e., 15) evaluated criterion-concurrent validity than any other type of validity via relations to other variables. A foundational goal of population health research is to evaluate predictors (i.e., health determinants, risk or protective factors) of health outcomes (Kindig and Stoddart 2003). The term predictor implies causality – social support causes, at least partially, the health

outcome under investigation. Thus, even if research is cross-sectional, temporality is presumed. Interestingly, there were no studies of criterion-predictive validity – the validity evidence that most directly evaluates if a score predicts, over time, a theoretically expected criterion (i.e., outcome). Demonstrating predictive validity would add credibility to the argument that a predictor is causing a certain outcome, and argument central to population health research.

Conclusion

As noted above, validity is dependent upon context, which includes the populations of interest and purposes of score interpretation (Messick 1995). In population health, measures are used to evaluate health determinants and outcomes within and between populations. Thus, we encourage population health researchers to evaluate validity in the context of the populations they are investigating – evaluating validity in entire populations as well as within and between subpopulations. As an example, let us say researchers are interested in evaluating the validity of the MOS-SS in the U.S. population. First, they would examine validity in the context of a representative sample of the U.S. population. Next, they would identify subpopulations of interest to investigate validity both between and within those populations. We recognize that this is a potential *rabbit hole* insofar as there are an almost limitless number of subpopulations. Thus, we encourage researchers to consider (a) what subpopulations are of most interest to them and (b) what subpopulations would past research and theory indicate may differentially respond on the measure(s) of interest. For example, in some populations, religious affiliation may greatly vary and influence the construct of interest. Thus, validity evidence within and between religions should be evaluated. However, within other populations, religious affiliation or lack thereof is relatively homogeneous. Thus, how religious affiliation impacts the validity of the measure's score is of little interest or value. To further illustrate our thinking, we have presented a figure that depicts how population health researchers could divide populations and subpopulations in which to investigate validity (see Fig. 13.1). While this may seem daunting, we want to reiterate that the purpose of population health research is to evaluate the relation of health determinants and health outcomes between and within populations. Therefore, if validity evidence is not evaluated in different populations, both null and significant findings in population health research may not be valid.

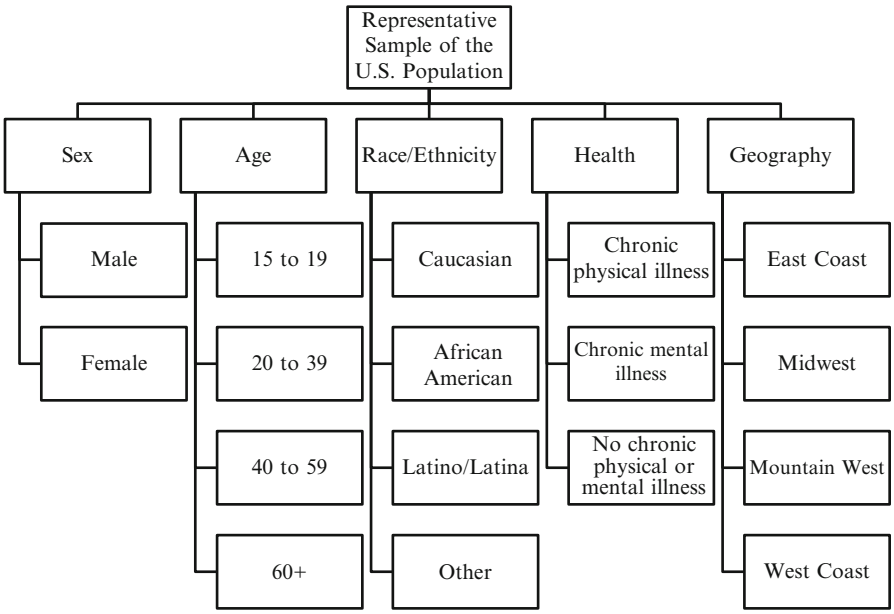

Fig. 13.1 Example of how a population can be divided into meaningful subpopulations for between- and within-group validation study

References

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Forer, B., & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research, 103*(2), 231–265. doi:10.1007/s11205-011-9844-3.

Kindig, D., & Stoddart, G. (2003). What is population health? *American Journal of Public Health, 93*(3), 380–383. doi:10.2105/AJPH.93.3.380.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. doi:10.1037/0003-066X.50.9.741.

Moser, A., Stuck, A. E., Silliman, R. A., Ganz, P. A., & Clough-Gorr, K. M. (2012). The eight-item modified Medical Outcomes Study Social Support Survey: Psychometric evaluation showed excellent performance. *Journal of Clinical Epidemiology, 65*(10), 1107–1116. doi:10.1016/j.jclinepi.2012.04.007.

Robitaille, A., Orpana, H., & McIntosh, C. N. (2011). Psychometric properties, factorial structure, and measurement invariance of the English and French versions of the Medical Outcomes Study Social Support Scale. *Health Reports, 22*(2), 33–40.

Sherbourne, C. D., & Stewart, A. L. (1991). The MOS social support survey. *Social Science & Medicine, 32*(6), 705–714. doi:10.1016/0277-9536(91)90150-B.

Statistics Canada. (2011). *CCHS annual component 2010 and 2009–2010 user guide in English*. Ottawa, ON. Retrieved from http://abacus.library.ubc.ca.ezproxy.library.ubc.ca/bitstream/10573/41532/11/guide_e.pdf

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. doi:10.1177/109442810031002.

Zumbo, B. D., & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J. A. Bovaird & K. F. Geisinger (Eds.), *High stakes testing in education: Science and practice in K-12 settings* (pp. 177–190). Washington, DC: American Psychological Association.

Chapter 14
Medical Outcomes Study Short Form-36 (SF-36) and the World Health Organization Quality of Life (WHOQoL) Assessment: Reporting of Psychometric Validity Evidence

Eric K.H. Chan, Bruno D. Zumbo, Wen Zhang, Michelle Y. Chen, Ira Darmawanti, and Olievia P. Mulyana

Psychometric patient-reported outcome (PRO) instruments are increasingly used to accompany traditional biomedical measures to evaluate health outcomes. In the guidelines *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims* by the Food and Drug Administration (FDA 2009), PRO is defined 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). Most PRO instruments are multidimensional and assess various domains of health status and quality of life, such as general health, well-being, physical, emotional, cognitive, social, and sexual functioning, as well as symptoms, side effects, and toxicity (Fayers and Machin 2007). PRO instruments help researchers, clinicians, and policy makers understand from a patient's points of view whether medical and healthcare interventions are effective. The FDA now takes PRO data into consideration in the appraisal of health technologies. Task forces and guidelines on the reporting of validity evidence for psychometric PRO instruments in clinical trials have also been established (Brundage et al. 2013; Calvert et al. 2013).

According to the Patient-Reported Outcome and Quality of Life Instruments Database (PROQoLID), over 700 PRO instruments exist. Validity is a fundamental issue in the evaluation and development of psychometric PRO instruments. The theories of validity and methods for validation have become more advanced, expansive, and complex during the past few decades. In a seminal paper on the

E.K.H. Chan, Ph.D. (✉) • B.D. Zumbo, Ph.D. • W. Zhang • M.Y. Chen
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; bruno.zumbo@ubc.ca; zwlisa@gmail.com; michellec2004@gmail.com

I. Darmawanti • O.P. Mulyana
Department of Educational Psychology and Guidance, State University of Surabaya, Ketintang Baru XIV/2, Surabaya, East Java 60231, Indonesia
e-mail: ira.darmawanti@gmail.com; olimulya@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_14,

unitary view of validity by Messick (1989), validity is defined as "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). This definition suggests that there is no singular source of evidence sufficient to support a validity claim. Messick's work is highly influential and his view is reflected in the most current edition of the *Standards for Educational and Psychological Testing* (AERA et al. 1999), which suggests that five sources of validity evidence are needed to support PRO score inferences. The five sources include (1) test content, (2) internal structure (the extent to which items relate to each other and represent the construct of interest), (3) associations with other variables (e.g., convergent, discriminant, concurrent, and predictive validity), (4) response processes (the cognitive processes involved in item responses), and (5) consequences (intended use and misuse).

Although Messick's (1989) view of validity has been published for over 20 years, it does not appear to have been adopted in the empirical literature and hence certain sources of validity evidence are neglected in current validation practices. For example, in a study of an internal medicine student assessment system, the reporting of response processes and consequences were not present (Auewarakul et al. 2005). Similar results were obtained in studies in education and psychology (Cizek et al. 2008, 2010).

The purpose of the present study was to examine the validity evidence reported in the academic literature for the Medical Outcomes Study Short Form-36 (SF-36) and the World Health Organization Quality of Life (WHOQoL) Assessment, two widely used PRO instruments for assessing generic health status and quality of life in health (as opposed to disease-specific instruments which focus on a particular disease conditions). Our focus was not on the quality appraisal of the two instruments, but rather on informing and improving the validation practices of the two instruments and to examine the extent to which the validation practices of the two instruments align with the contemporary validity perspectives. The following research question was addressed: What sources of validity evidence are reported for the SF-36 and WHOQoL?

SF-36

Developed by John Ware and his colleagues (1993, 1994), the SF-36 is a 36-item self-report instrument developed in the 1980s to assess generic health status in individuals aged 14 or older primarily for research and population health monitoring purposes. The 36 items cover eight domains scaled from 0 to 100, with higher values indicating better health status. The eight domains include: physical function (PF), role physical (RP), bodily pain (BP), general health (GH), vitality (VI), social functioning (SF), role emotional (RE), and mental health (MH). The eight domains form two higher-ordered clusters, namely (1) general physical health and (2) general mental health. The PF, RP, BP, and GH domains form the general physical health

cluster and the MH, RE, SF, and VT domains form the general mental health cluster. For reporting purposes, each domain is transformed into standard scores, with a mean of 50 and standard deviation of 10. The instrument can be self-administered, computer-administered, or administered by a trained interviewer in person or via telephone. It takes between 5 and 10 min to complete.

The SF-36 scores possess internal consistency and test-retest reliability evidence, as well as content, concurrent, criterion, construct, predictive validity, and internal structure (factor analysis) evidence (Ware et al. 1993, 1994). The instrument has been included in general population health surveys in many countries, as well as patients of different age groups with different diseases (Ware et al. 1995). The SF-36 has been translated into a number of languages and is widely used globally.

World Health Organization Quality of Life Assessment

The WHOQoL Assessment project began in 1991 at the Division of Mental Health of the World Health Organization (WHO). The WHO defines quality of life as individuals' perception of their position in life in the context of the culture and value systems they live in and in relation to their goals, expectations, standards, and concerns (The WHOQoL Group 1994). This definition suggests that quality of life is a broad concept, and can be affected by an individual's physical health, psychological state, personal beliefs, social relationships and their relationship to salient features of their environment in a complex way. The focus of the WHOQoL is on individuals' views of their own well-being, which differs from many medical assessments obtained by health workers' examinations and laboratory tests. It not only inquires about the functioning of patients across a range of areas but also how satisfied the patients are with their functioning and with effects of treatment.

The items of the WHOQoL were written based on the statements made by patients of a wide range of diseases, healthy people, and health professionals in different cultures. The WHOQoL covers 6 broad domains of quality of life, and 24 facets (see Table 14.1). Four items are included for each facet, as well as four general items covering subjective overall quality of life and health, producing a total of 100 items in the assessment. All items are rated on a five-point scale. The core WHOQoL instrument assesses quality of life in a variety of situations and populations. In addition, modules are being developed to allow more detailed assessments of specific populations (e.g. cancer patients, refugees, elderly, and those with life-threatening diseases, such as HIV/AIDS). The WHOQoL is also found to be cross-culturally valid and sensitive. The WHOQoL is now available in over 20 different languages and continues to be translated into additional languages.

Table 14.1 The structure of the WHOQOL

Domains	Facets incorporated within domains
Overall quality of life and general health	
(A) Physical health	Energy and fatigue
	Pain and discomfort
	Sleep and rest
(B) Psychological	Bodily image and appearance
	Negative feelings
	Positive feelings
	Self-esteem
	Thinking, learning, memory and concentration
(C) Level of independence	Mobility
	Activities of daily living
	Dependence on medicinal substances and medical aids
	Work capacity
(D) Social relations	Personal relationships
	Social support
	Sexual activity
(E) Environment	Financial resources
	Freedom, physical safety and security
	Health and social care: accessibility and quality
	Home environment
	Opportunities for acquiring new information and skills
	Participation in and opportunities for recreation/leisure
	Physical environment (pollution/noise/traffic/climate)
	Transport
(F) Spirituality/Religion/Personal beliefs	Religion/Spirituality/Personal beliefs (Single facet)

Methods

Database Search

We conducted a systematic search in January 2013 on PubMed, PsycINFO, Cinahl, and Embase. For the SF-36, the following search keywords were used: "SF-36" AND "development" OR "measurement" OR "psychometric" OR "psychometrics" OR "reliable" OR "reliability" OR "valid" OR "validation" OR "validity". For the WHOQoL, the following keywords were used: "World Health Organization Quality of Life Assessment" OR "WHOQOL-100" OR "WHOQOL-BREF" AND "development" OR "measurement" OR "psychometric" OR "psychometrics" OR "reliable" OR "reliability" OR "valid" OR "validation" OR "validity". The searches were limited to TITLE for both the SF-36 and WHOQOL. After duplicates were removed the search resulted in 764 articles in total for the SF-36 and 384 articles in total for the WHOQoL.

Screening

To be included in this review, each study must (1) explicitly state that validity is the focus/objective and (2) be empirical studies. We excluded (1) opinion papers and editorials, (2) reviews, systematic reviews, and meta-analyses, (3) guidelines, task force papers, recommendations, and statistical applications, (4) conference proceedings/abstracts, and (5) utility, econometric, preference-based, and other non-validation studies. The present review was delimited to including studies using the psychometric approach to validation, because studies using the economic or utility approach come from a different tradition of how one develops and "validates" instruments, and the language and framework are different from the psychometric approach (Kopec and Willison 2003; Richardson and Zumbo 2000).

Although retrieving and reviewing all validity and validation articles on the SF-36 and would be ideal, due to the large number of articles and our limited resources, our team chose to randomly select and include 30 empirical articles (15 each of the SF-36 and WHOQoL) in the present study.

Coding

A coding form was developed to record the characteristics and validity evidence presented in each of the 30 selected SF-36 and WHOQoL articles. Following the modern validity framework as stated in the most current version of the *Standards for Educational and Psychological Testing* (AERA et al. 1999) and building from previous published studies (e.g., Cizek et al. 2008, 2010), our team coded the following sources of validity evidence: face, content, construct, predictive, concurrent, convergent, discriminant, response processes, consequences, and other. Our coding was based on what the authors of the articles reported and the methodological procedures employed. For instance, if "discriminant validity" is explicitly stated in an article, discriminant validity was coded. If the "think-aloud" procedure was employed to investigate people's responses yet the evidence was not explicitly stated as "response processes", it was still coded as response process evidence. Similarly, if factor analytic results were reported but were not explicitly stated as internal structure evidence in the paper, we still coded them as internal structure evidence.

In this study, four of the authors completed the coding. We randomly selected three SF-36 and three WHOQoL articles (from the 30) and the six articles were coded by all four individuals. The remaining 24 articles were equally divided, therefore, each of the four individuals individually coded six (three SF-36 and three WHOQOL) of the 24 articles. In total, each of the four individuals coded 12 (six unique [three SF-36 and three WHOQOL] and six common [three SF-36 and three WHOQOL]) articles.

In terms of the agreement among raters we report the details for the SF-36 because the WHOQOL findings were very similar. Disagreement among the four raters occurred in 5 of the 30 multiple ratings (i.e., 3 articles for the 10 sources of validity that were coded); and for the content, predictive and convergent evidence codes only one rater disagreed with the other three. Internal structure was a bit more complicated because two raters disagreed on two of the articles which reported unconventional psychometric methods of internal structure. There was no disagreement among the raters for the face, construct, concurrent, discriminant, consequences, and response process codes. Overall, we consider this a high consistency among our four raters. In terms of the final coding for reporting validation practices, any disagreements in the coding results were reviewed and inconsistencies were resolved by the first author.

Results

SF-36

Our findings show that a broad perspective on the possible sources of validity evidence is reported in the published literature for the SF-36. Researchers conducting validation studies to support the SF-36 score inferences are not relying on only one sources of validity evidence at the exclusion of all other sources. As shown in Table 14.2, the number of sources of validity evidence reported per study ranged from 0 to 5, with a mode of one. Two studies reported zero source of validity yet presented themselves as validation studies, and one study reported five sources of validity evidence.

As seen in Table 14.3, internal structure and construct validity, each reported in close to half of the papers included in the present study, were the most frequently reported sources of validity evidence to support the interpretation of the SF-36 scores. Examples of the reported statistical methods to examine construct validity evidence included factor analysis, correlations with other instruments, convergent and discriminant evidence, and analysis of group differences.

About one third of the studies reported convergent validity evidence. Discriminant validity, which serves as a baseline to compare convergent validity, was reported in one fifth of the papers. About 13 % of the papers reported concurrent validity evidence. One study presented predictive validity evidence. No study reported evidence on face validity, content validity, or response processes. Consequences, which are emerging as an important source of validity evidence, were also not reported in any of the 15 SF-36 studies included in the present analysis (see Table 14.3).

Table 14.2 Frequency of number of validity sources reported for SF-36

Number of sources	Frequency	Percent
0	5	33.3
1	3	20.0
2	3	20.0
3	2	13.3
4	1	6.7
5	1	6.7
Total	**15**	**100**

Table 14.3 Sources of validity reported for SF-36[a]

Source of validity	Number of papers (n = 15)	Percent of papers
Construct	7	46.7
Internal structure	7	46.7
Convergent	5	33.3
Discriminant	3	20.0
Concurrent	2	13.3
Predictive	1	6.7
Content	0	0
Face	0	0
Response processes	0	0
Consequences	0	0

[a]A paper can report more than one source of validity

Table 14.4 Frequency of number of validity sources reported for WHOQoL

Number of sources	Frequency	Percent
0	3	20.0
1	1	6.7
2	1	6.7
3	3	20.0
4	6	40.0
5	1	6.7
Total	**15**	**100**

WHOQoL

Similar to the SF-36 results, a broad perspective on the possible sources of validity evidence is reported for the WHOQoL. Researchers conducting validation studies to support the WHOQoL score inferences are not relying on only one sources of validity evidence in the exclusion of all other sources. As shown in Table 14.4, the number of sources of validity evidence reported per study ranged from 0 to 5, with a mode of four. Three studies that presented themselves as validation studies had zero source of validity evidence and two studies reported five sources of validity evidence. From Table 14.5 one can see that discriminant validity was found to be the most frequently reported source of evidence, reported in two thirds of the papers.

Table 14.5 Sources of validity reported for WHOQoL[a]

Source of validity	Number of papers (n = 15)	Percent of papers
Discriminant	10	66.7
Internal structure	9	60.0
Construct	8	53.3
Convergent	7	46.7
Content	4	26.7
Predictive	1	6.7
Concurrent	1	6.7
Face	0	0
Response processes	0	0
Consequences	0	0

[a]A paper can report more than one source of validity

Over half of the studies reported internal structure or construct validity evidence. Confirmatory factor analysis, correlations, and convergent and discriminant evidence were the reported analytic methods to investigate construct validity of the WHOQoL scores. Convergent validity was reported in nearly half of the papers. About a quarter of the papers reported content validity evidence. Examples of methods to examine content validity included the use of (1) proportion of substantive agreement and substantive validity coefficient, (2) item-domain correlation, and (3) factor analysis which these authors considered as a statistical method for examining content validity. Concurrent and predictive validity evidence were not commonly reported (each was only reported in one [6.7 %] of the studies). Face validity evidence was never reported. Response processes and consequences, which are emerging as an important source of validity evidence, were not reported in any of the 15 studies (see Table 14.5).

Discussion

Validity theories and validation methodology have grown extensively during the past several decades. Examining the reporting of validity evidence published in academic journals is a good way to understand the current practices of validation and to generate recommendations to advance the field of validity theory and validation practices. The purpose of this study was to review the reporting of validity evidence for the SF-36 and WHOQoL, two widely used psychometric PRO instruments.

Our results revealed that the sources of validity evidence reported for the SF-36 and WHOQoL vary and hence researchers are not focusing on one source of validity evidence at the exclusion of all others to support the score interpretation of the two instruments. Internal structure and construct validity were the two most widely reported sources of evidence for both the SF-36 and WHOQoL. The other two commonly reported sources of validity for the two instruments were convergent and discriminant.

Although all sources of validity need not be used in every case of validation, the sources of validity evidence should depend on the purpose of the instrument and the particular population for which it is intended to be used. It is interesting to note that response processes and consequences are never reported in our samples for the SF-36 and WHOQoL. These two sources are important in the process of validation and as proponents of their inclusion note, failing to include these two important sources of validity evidence to support the score interpretation of the SF-36 and WHOQoL could affect the quality of PRO research and may in turn influence the quality of healthcare that patients receive (Hubley and Zumbo 2011, 2013; Messick 1989, 1995; Zumbo 2007, 2009).

A few points are noteworthy regarding response processes and consequences. In the investigation of response processes, the focus is on the cognitive or thinking processes involved when patients respond to items on a PRO instrument. As Zumbo (2009) notes, having an understanding of how and why individuals respond to items the way that they do goes a long way to understanding what one is measuring with an instrument. Response processes also concerns the influence individual characteristics have on their responses to PRO items. For example, when asked about social role functioning, it is important that people are reasoning about their social role functioning and their responses should not be influenced by characteristics such as social norm or social expectations. Our findings reveal that this source of validity evidence was rarely reported for the SF-36 and WHOQoL. Given the importance of examining the substantive aspect of validity has been stressed (AERA et al. 1999; Messick 1989, 1995) and is emerging as central to validity claims (Hubley and Zumbo 2011, 2013; Messick 1995; Zumbo 2007, 2009), more research is needed to investigate the response processes involved in responding to the items on the SF-36 and WHOQoL.

Consequences, which include both positive and negative consequences of intended use of PRO scores, are also emerging as central to validity claims (AERA et al. 1999; Hubley and Zumbo 2011, 2013; Messick 1989, 1995; Zumbo 2007, 2009). However consequences were not reported in any of the 30 papers on the validity of the SF-36 and WHOQoL. First, intended use is the claims or decisions we want to make based on the scores on a PRO instrument and is part of the entire validation activities. Issues such as construct underrepresentation and irrelevant variance can negatively influence the intended use of a PRO instrument. As Messick highlights in the context of general measurement theory, the misuse of PRO scores does not invalidate their appropriate use. That is, consequences are not about misuse but rather the correct use of an instrument. This is a subtle but important point more fully explicated in Hubley and Zumbo (2011).

Several limitations of this study need to be discussed. First, the present study is limited by the random selection of a small number of validity papers on the SF-36 and WHOQoL. Although it would be ideal to include all studies, due to our limited resources we were unable to do so and our results may not be generalizable to all of the validation studies conducted on the two instruments. Full systematic reviews are needed in the future.

Second, our findings do not imply that researchers are not aware of sources of validity such as response processes and consequences, nor do our findings point to the fact that researchers conducting validation research only focus on a limited number of validity sources. It is possible that researchers conducting validation research to support the score interpretation of the SF-36 and WHOQoL did in fact follow the modern view of validity and investigate all sources of validity evidence (including response processes and consequence) but due to reasons such as limited journal space, decided not to report all sources of validity evidence in their papers. The design of the present study did not allow us to examine this issue.

In conclusion, the findings of the present study reveal that the modern view of validity is not reflected in the validation practices to support the score interpretations of the SF-36 and WHOQoL. Perhaps explicit recommendations need to be outlined to ensure the reporting of validity evidence for PRO instruments covers the different sources of validity to support valid score interpretations and healthcare decision making.

References[1]

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.

Auewarakul, C., Downing, S. M., Jaturatamrong, U., & Praditsuwan, R. (2005). Sources of validity evidence for an internal medicine student evaluation system: An evaluative study of assessment methods. *Medical Education, 39*, 276–283.

*Augustovski, F. A., Lewin, G., Elorrio, E. G., & Rubinstein, A. (2008). The Argentine-Spanish SF-36 Health Survey was successfully validated for local outcome research. *Journal of Clinical Epidemiology, 61*, 1279–1284.

*Berlim, M. T., Pavanello, D. P., Caldieraro, M. A. K., & Fleck, M. P. A. (2005). Reliability and validity of the WHOQOL BREF in a sample of Brazilian outpatients with major depression. *Quality of Life Research, 14*, 561–564.

*Bonomi, A. E., Patrick, D. L., Bushnell, D. M., & Martin, M. (2000). Validation of the United States' version of the World Health Organization Quality of Life (WHOQOL) instrument. *Journal of Clinical Epidemiology, 53*, 1–12.

Brundage, M., Blazeby, J., Revicki, D., Bass, B., de Vet, H., Duffy, H., et al. (2013). Patient-reported outcomes in randomized clinical trials: Development of ISOQOL reporting standards. *Quality of Life Research, 22*, 1161–1175.

Calvert, M., Blazeby, J., Altma, D. G., Revicki, D. A., Moher, D., & Brundage, M. D., for the CONSORT PRO Group. (2013). Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. *Journal of the American Medical Association, 309*, 814–822.

*Chien, C.-W., Wang, J.-D., Yao, G., Sheu, C.-F., & Hsieh, C.-L. (2007). Development and validation of a WHOQOL-BREF Taiwanese audio player-assisted interview version for the elderly who use a spoken dialect. *Quality of Life Research, 16*, 1375–1381.

[1] References marked with an asterisk indicate studies included in the review.

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.

*Colbourn, T., Masache, G., & Skordis-Worrall, J. (2012). Development, reliability and validity of the Chichewa WHOQOL-BREF in adults in Lilongwe, Malawi. *BioMed Central Research Notes, 5*, 346.

*Dallmeijer, A. J., de Groot, V., Roorda, L. D., Schepers, V. P. M., Lindeman, E., van den Berg, L. H., Beelen, A., & Dekker, J. (2007). Cross-diagnostic validity of the SF-36 physical functioning scale in patients with stroke, multiple sclerosis and amyotrophic lateral sclerosis: A study using Rasch analysis. *Journal of Rehabilitation Medicine, 39*, 163–169.

*Edgar, D., Dawson, A., Hankey, G., Phillips, M., & Wood, F. (2010). Demonstration of the validity of the SF-36 for measurement of the temporal recovery of quality of life outcomes in burns survivors. *Burns, 36*, 1013–1020.

Fayers, P., & Machin, D. (2007). *Quality of life: The assessment, analysis and interpretation of patient-reported outcomes* (2nd ed.). Chichester: Wiley.

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.

*Garcia-Rea, E. A., & LePage, J. P. (2010). Reliability and validity of the World Health Organization Quality of Life: Brief version (WHOQOL-BREF) in a homeless substance dependent veteran population. *Social Indicators Research, 99*, 333–340.

*Hsiung, P. C., Fang, C. T., Wu, C. H., Sheng, W. H., Chen, S. C., Wang, J. D., & Yao, G. (2011). Validation of the WHOQOL-HIV BREF among HIV-infected patients in Taiwan. *AIDS Care, 23*, 1035–1042.

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.

*Jahanlou, A. S., & Karami, N. A. (2011). WHO quality of life-BREF 26 questionnaire: Reliability and validity of the Persian version and compare it with Iranian diabetics quality of life questionnaire in diabetic patients. *Primary Care Diabetes, 5*, 103–107.

*Keller, S. D., Ware, J. E., Jr., Hatoum, H. T., & Kong, S. X. (1999). The SF-36 Arthritis-Specific Health Index (ASHI): II. Tests of validity in four clinical trials. *Medical Care, 37*, 51–60.

Kopec, J. A., & Willison, K. D. (2003). A comparative review of four preference-weighted measures of health-related quality of life. *Journal of Clinical Epidemiology, 56*, 317–325.

*Lam, C. L., Tse, E. Y., Gandek, B., & Fong, D. Y. (2005). The SF-36 summary scales were valid, reliable, and equivalent in a Chinese population. *Journal of Clinical Epidemiology, 58*, 815–822.

*Lera, L., Fuentes-García, A., Sánchez, H., & Albala, C. (2013). Validity and reliability of the sf-36 in Chilean older adults: The ALEXANDROS study. *European Journal of Ageing, 10*, 127–134.

*Lucas-Carrasco, R., Skevington, S. M., Gómez-Benito, J., Rejas, J., & March, J. (2011). Using the WHOQOL-BREF in persons with dementia: A validation study. *Alzheimer Disease and Associated Disorders, 25*, 345–351.

*Masthoff, E. D., Trompenaars, F. J., Van Heck, G. L., Hodiamont, P. P., & De Vries, J. (2005). Validation of the WHO Quality of Life assessment instrument (WHOQOL-100) in a population of Dutch adult psychiatric outpatients. *European Psychiatry, 20*, 465–473.

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

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

*Montazeri, A., Goshtasebi, A., Vahdaninia, M., & Gandek, B. (2005). The Short Form Health Survey (SF-36): Translation and validation study of the Iranian version. *Quality of Life Research, 14*, 875–882.

*Osborne, R. H., Hawthorne, G., Lew, E. A., & Gray, L. C. (2003). Quality of life assessment in the community-dwelling elderly: Validation of the Assessment of Quality of Life (AQoL) instrument and comparison with the SF-36. *Journal of Clinical Epidemiology, 56*, 138–147.

Richardson, C. G., & Zumbo, B. D. (2000). A statistical examination of the Health Utility Index-Mark III as a summary measure of health. *Social Indicators Research, 51*, 171–191.

*Rotstein, Z., Barak, Y., Noy, S., & Achiron, A. (2000). Quality of life in multiple sclerosis: Development and validation of the 'RAYS' scale and comparison with the SF-36. *International Journal for Quality in Health Care, 12*, 511–517.

*Saddki, N., Noor, M. M., Norbanee, T. H., Rusli, M. A., Norzila, Z., Zaharah, S., Sarimah, A., Norsarwany, M., Asrenee, A. R., & Zarina, Z. A. (2009). Validity and reliability of the Malay version of WHOQOL-HIV BREF in patients with HIV infection. *AIDS Care, 21*, 1271–1278.

*Scott, K. M., Sarfati, D., Tobias, M. I., & Haslett, S. J. (2000). A challenge to the cross-cultural validity of the SF-36 health survey: Factor structure in Māori, Pacific and New Zealand European ethnic groups. *Social Science & Medicine, 51*, 1655–1664.

*Seymour, D. G., Ball, A. E., Russell, E. M., Primrose, W. R., Garratt, A. M., & Crawford, J. R. (2001). Problems in using health survey questionnaires in older patients with physical disabilities. The reliability and validity of the SF-36 and the effect of cognitive impairment. *Journal of Evaluation in Clinical Practice, 7*, 411–418.

*Skevington, S. M., & Wright, A. (2001). Changes in the quality of life of patients receiving antidepressant medication in primary care: Validation of the WHOQOL-100. *British Journal of Psychiatry, 178*, 261–267.

*SooHoo, N. F., McDonald, A. P., Seiler, I. J. G., & McGillivary, G. R. (2002). Evaluation of the construct validity of the DASH questionnaire by correlation to the SF-36. *Journal of Hand Surgery, 27*, 537–541.

*Takeshita, K., Maruyama, T., Matsudaira, K., Murakami, M., Higashikawa, A., & Nakamura, K. (2006). Validity and reliability of SRSI and SF-36 in Japanese patients with scoliosis. *Studies in Health Technology and Informatics, 123*, 337–342.

The WHOQoL Group. (1994). The development of the World Health Organization Quality of Life Assessment Instrument (the WHOQoL). In J. Orley & W. Kuyken (Eds.), *Quality of life assessment: International perspectives*. Heidelberg: Springer.

*Trompenaars, F. J., Masthoff, E. D., Van Heck, G. L., Hodiamont, P. P., & De Vries, J. (2005). Content validity, construct validity, and reliability of the WHOQOL-Bref in a population of Dutch adult psychiatric outpatients. *Quality of Life Research, 14*, 151–160.

*Tsutsumi, A., Izutsu, T., Kato, S., Islam, M. A., Yamada, H. S., Kato, H., & Wakai, S. (2006). Reliability and validity of the Bangla version of WHOQOL-BREF in an adult population in Dhaka, Bangladesh. *Psychiatry and Clinical Neurosciences, 60*, 493–498.

*Van Leeuwen, C. M. C., Van Der Woude, L. H. V., & Post, M. W. M. (2012). Validity of the mental health subscale of the SF-36 in persons with spinal cord injury. *Spinal Cord, 50*, 707–710.

Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 health survey manual and interpretation guide*. Boston: New England Medical Center, The Health Institute.

Ware, J. E., Kosinski, M., & Keller, S. K. (1994). *SF-36 physical and mental health summary scales: A user's manual*. Boston: The Health Institute.

Ware, J. E., Keller, S. D., Gandek, B., Brazier, J. E., & Sullivan, M. (1995). Evaluating translations of health status questionnaires: Methods from the IQOLA project. *International Journal of Technology Assessment in Health Care, 11*, 525–551.

*Webster, J., Nicholas, C., Valacott, C., Cridland, N., & Fawcett, L. (2010). Validation of the WHOQOL-BREF among women following childbirth. *The Australian & New Zealand Journal of Obstetrics & Gynaecology, 50*, 132–137.

*Wyss, K., Wagner, A. K., Whiting, D., Mtasiwa, D. M., Tanner, M., Gandek, B., & Kilima, P. M. (1999). Validation of the Kiswahili version of the SF-36 Health Survey in a representative sample of an urban population in Tanzania. *Quality of Life Research, 8*, 111–120.

*Yao, G., Wu, C.-h., & Yang, C.-t. (2008). Examining the content validity of the WHOQOL-BREF from respondents' perspective by quantitative methods. *Social Indicators Research, 85*, 483–498.

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/ Boston: 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: Information Age Publishing.

Chapter 15
Reporting of Validity Evidence in the Field of Health Care: A Focus on Papers Published in *Value in Health*

Eric K.H. Chan, Bruno D. Zumbo, Ira Darmawanti, and Olievia P. Mulyana

Health care research is, in broad terms, meant to guide policy and decision makers in considering alternative treatments, evaluating treatment effectiveness, health services evaluation, and health care resource allocation. Psychometric instruments based on self-report, or ratings by others, are increasingly used in health care to compliment pharmacoeconomics and outcomes research. For instance, more emphases have been placed on the use of patient-reported outcomes (PRO), and particularly health-related quality of life and wellbeing, because patient perspectives are unique, are central components in diagnosis and treatment, and can complement traditional biomedical indicators of disease status and treatment effectiveness (Acquadro et al. 2003). Other areas in health care research, such as the assessment of physician psychological attributes (Hojat 2007; Hojat et al. 2001) and clinical competency (Auewarakul et al. 2005) also utilize psychometric instruments. Therefore, the use of psychometric instruments has far-reaching consequences in health care.

Validity is pivotally important in the development and evaluation of psychometric instruments (AERA et al. 1999; Messick 1989), including instruments used in health care. For instance, in the recently published industry guidance titled "Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims" (US Food and Drug Administration 2009), the Food and Drug Administration (FDA) discussed validity issues. Other groups such as the Scientific Advisory Committee of the Medical Outcomes Trust (2002) and the joint Pharmaceutical Research and Manufacturers of America Health Outcomes

E.K.H. Chan (✉) • B.D. Zumbo, Ph.D.
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; bruno.zumbo@ubc.ca

I. Darmawanti • O.P. Mulyana
Department of Educational Psychology and Guidance, State University of Surabaya, Ketintang Baru XIV/2, Surabaya, East Java 60231, Indonesia

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_15,

Committee (PhRMA HOC) and the Division of Drug, Marketing, Advertising, and Communications of the FDA (DDMAC FDA) (Santanello et al. 2002) have also published articles discussing the importance of validity for psychometric instruments in health care.

Validity theory and validation methods have become more complex and expansive over the past several decades. There is an agreement among validity experts that the accumulation and integration of evidence from different sources is needed to support the validity of the interpretation and inferences made from the scores arising from these measures (AERA et al. 1999; Kane 2006; Messick 1989; Zumbo 2007, 2009). The contemporary view of validity contends that in addition to the traditional sources of validity such as content, relations to other variables (e.g., discriminant, and convergent validity), and internal structure, 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 validating psychometric instruments (AERA et al. 1999; Messick 1989, 1995; Hubley and Zumbo 2011).

A good way to investigate how a psychometric validation study is designed is by examining the reporting characteristics. For instance, although not studies of psychometric validation practices, studies have investigated the reporting of methodological and statistical details in randomized controlled trials (Chan and Altman 2005) and systematic reviews (Moher et al. 2007). With respect to psychometric validity, studies examining the reporting of validity evidence in the psychology and education literature have shown that a number of sources of validity evidence are not presented, with only 1.8 % and 2.5 % of the studies reporting response processes and consequences respectively (Cizek et al. 2008, 2010). Similarly, a review of clinical assessment in internal medicine has found that the reporting of response processes and consequences were absent (Auewarakul et al. 2005).

With an aim towards investigating the validity evidence and refining and improving the practice of psychometric validation in health care, the purpose of the present study was to investigate the reporting of validity evidence in papers published in *Value in Health*, the official journal of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR). This scholarly society, and its official research journal, was selected because, as noted in the society's mission statement, ISPOR is recognized globally as the authority for outcomes research and its use in health care decisions towards improved health. As such, the papers published in that journal are authoritative resources for shaping health care practices and research and serve as a fruitful ground from which to investigate psychometric validation practices in health care. Our research question was: What sources of validity evidences are reported in the validation of psychometric instruments published in the journal? Our focus was on informing validation practice, not on evaluating the quality of the psychometric instruments.

Methods

A systematic search using the official website of the journal was conducted in December 2010. We searched for papers published since the journal's inception (January 1998) to December 2010. We searched both the titles and abstracts. Keywords used in the search included "*development* OR *measurement* OR *psychometric* OR *psychometrics* OR *valid* OR *validation* OR *validity*." To be included in this review, each study must (1) be empirical psychometric studies and (2) be published between January 1998 and December 2010. We excluded (1) opinion papers and editorials, (2) reviews, systematic reviews, and meta-analyses, (3) guidelines, task force papers, recommendations, and statistical applications, (4) conference proceedings/abstracts, and (5) utility, econometric, preference-based, and other non-validation studies. We decided to exclude utility, preference-based, and related studies because these studies come from a different tradition of how one develops and "validates" instruments, and the language and framework are not the same as the psychometric approach (Kopec and Willison 2003; Richardson and Zumbo 2000). The present review was delimited to including studies using the psychometric approach to validation.

A coding sheet was developed to record the characteristics and validity evidence presented in each study. Building from earlier research (Cizek et al. 2008, 2010), variables included in our coding sheet were publication year and sources of validity evidence including face, content, construct, predictive, concurrent, convergent, discriminant, response processes, consequences, reliability, and other. We coded the sources of validity based on what the author(s) stated. For instance, if an author presented the correlation coefficient between two psychometric instruments but did not refer to, for example, criterion-related validity evidence, no validity evidence was coded. In other words, no subjective judgments were made as to the presence or the quality of the validity evidence. Each included article was double-coded independently by two of the authors, with an agreement of 88.1 %. Disagreements in the coding were discussed until consensus was reached.

Results and Conclusions

Search Process

Our search resulted in 347 abstracts and 126 titles. After initial screening (titles and abstracts), a total of 113 were retrieved and inclusion and exclusion applied. A final total of 68 papers were included in the present review. Of the instruments published in the journal, PRO measures accounted for the highest numbers. Others included an instrument designed for the evaluation of PRO measures (Valderas et al. 2008) and one that measures communications between physician and pharmacist from a

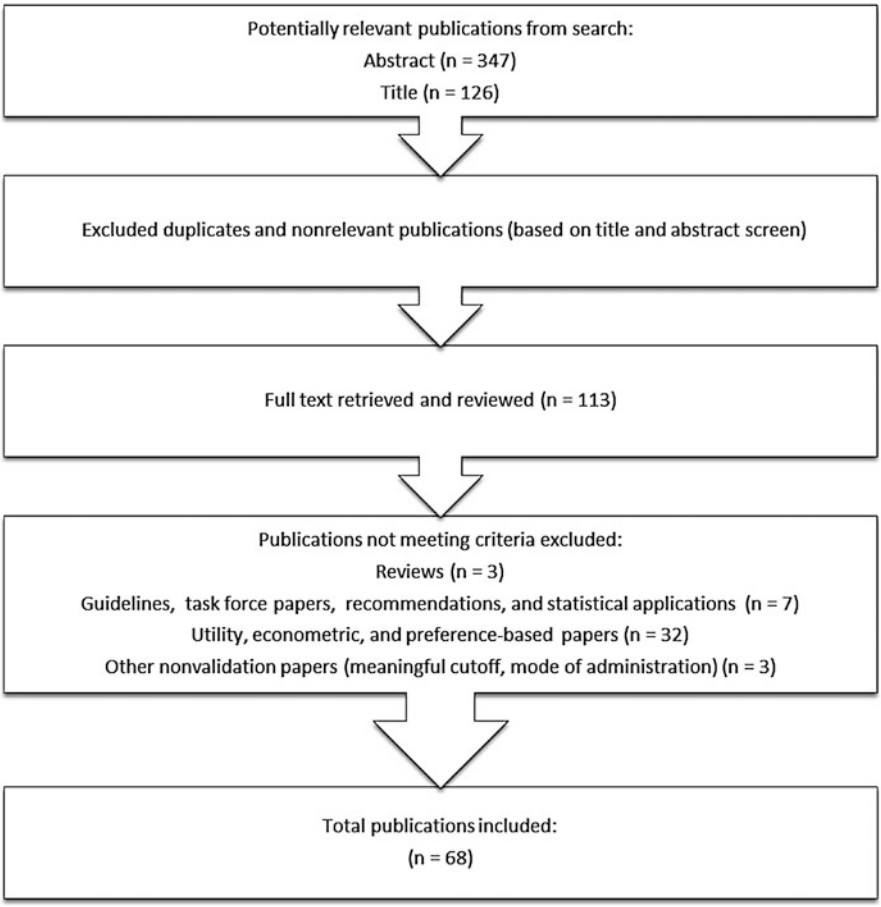

Fig. 15.1 Search process flowchart

physician's perspective (Zillich et al. 2005). Figure 15.1 presents the flowchart of the search process.

Descriptive Characteristics

Overall, there is an upward trend in the number of validity papers in the journal since its inception (see Table 15.1). When the journal began its publication in 1998, no study on validity was published. Less than 10 % of the validity studies were published between 1999 and 2004. Beginning in 2005, a higher number of validity studies were published each year, with a peak in 2009. Compared to 2009, there was a decrease in the number of validity studies in 2010.

Table 15.1 Year of publication

Year	Frequency	Percent
1998	0	0
1999	1	1.5
2000	1	1.5
2001	1	1.5
2002	2	2.9
2004	1	1.5
2005	8	11.8
2006	6	8.8
2007	4	5.9
2008	15	22.1
2009	18	26.5
2010	11	16.2
Total	**68**	**100**

Table 15.2 Frequency of number of validity sources reported

Number of sources	Frequency	Percent
0	6	8.8
1	11	16.2
2	10	14.7
3	14	20.6
4	17	25.0
5	7	10.3
6	3	4.4
Total	**68**	**100**

Reporting of Validity Evidence

Our findings revealed that the reporting of the sources of validity evidence in papers published in this journal varied. Researchers are not relying on only one source of validity evidence at the exclusion of all others and hence representing a broad perspective on the possible sources of validity evidence. As presented in Table 15.2, the number of sources of validity evidence reported per study ranged from zero to six, with a mode of four. A few studies had zero sources of evidence because the authors did not refer to any source of validity evidence although the papers were situated as validation studies. Internal consistency reliability was the most frequently reported source of evidence to support the consistency of the items and internal structure of an instrument, reported in over two thirds the papers. Half of the studies reported construct validity. Discriminant validity, which can serve as a baseline to compare convergent validity, is reported in one third of the papers. Similarly, one third of the papers reported evidence on convergent validity. There seems to be some confusion with the terminology in validity between discriminant and divergent validity, with a few of the studies using and reporting the term divergent validity as discriminant validity.

Table 15.3 Sources of validity reported[a]

Source of validity	Number	Percent
Internal consistency reliability	47	69.1
Construct	34	50.0
Convergent	23	33.8
Discriminant	23	33.8
Content	17	25.0
Known-Group	14	20.6
Criterion (concurrent or predictive)	14	20.6
Face	9	13.2
Response processes	3	4.4
Consequences	2	2.9

[a]A paper can report more than one source of validity

A quarter of the studies reported evidence on content validity. Evidence on "known-group validity", a term commonly used in the medical literature, was also reported in slightly over a fifth of the studies. Fourteen studies reported criterion validity evidence (13 of which reported only concurrent and the remaining one only reported predictive) and slightly over 10 % of the studies reported evidence on face validity. Evidence on predictive validity was only reported in one study. Response processes and consequences, which are important validity evidence, were also rarely reported (see Table 15.3).

Discussion

The purpose of this study was to review the reporting of validity evidence in papers published in *Value in Health*, with an eye towards informing future practice in health care. Authors of validity papers published in the journal are not focusing on one source of validity evidence at the exclusion of all other sources. Internal consistency and content type of validity was the most widely reported in the journal. Other commonly reported sources of validity include convergent and discriminant (including some articles referring it to divergent validity). Although the importance of response processes and consequences in validation have been well documented (Hubley and Zumbo 2011, 2013; Messick 1989, 1995; Zumbo 2007, 2009), these two sources are rarely presented in papers published in *Value in Health*. The absence of these two important sources of validity evidence could affect the medical care provided to patients.

Response processes were rarely reported. Although it is important to look at the substantive aspect of validity (AERA et al. 1999; Messick 1989, 1995), only about 4 % of the papers reported evidence related to response processes. Response processes are the thinking or cognitive processes involved when a patient responds to items on a health measure or when some performs a rating. In other words, the purpose is to investigate how and why people respond to questions or items the way they do. This sort of validity evidence is emerging as central to claims of psychometric validity (Hubley and Zumbo 2011, 2013; Messick 1995; Zumbo 2007, 2009).

Only two (2.9 %) papers mentioned consequences, commenting on the consequences and intended use of the instruments very briefly. Consequences in validity refer to the (1) intended use of measure scores and (2) misuse of measure scores (AERA et al. 1999; Hubley and Zumbo 2011, 2013; Messick 1989). Intended use concerns the decisions or claims one intends to make based on the scores on a psychometric instrument. It is part of the entire validation process and the intended use of an instrument can be influenced or weakened by issues such as construct underrepresentation or irrelevant variance. In depression for instance, males are consistently found have lower scores (i.e., less depressed) than their female counterparts. However, in a differential item functioning (DIF; Zumbo 1999) study on the Center for Epidemiologic Studies – Depression Scale (CES-D; Radloff 1977), several items were found to have gender DIF (Gelin and Zumbo 2003). Specifically, males were less likely to endorse several of the items (such as the item on "crying spells"), resulting in lower score among males. The lack of invariance in the depression scores between males and females may weaken the intended interpretation of the scores by confounding the interpretation with gender stereotypes and may have negative consequences on epidemiology findings, diagnostic decisions, and even insurance coverage.

Of equal importance in the concept of consequences is the issue of misuse of measure scores (Hubley and Zumbo 2011, 2013). For instance, clinicians cannot diagnose depression based on screening results. To give a diagnosis, additional clinical evaluation is needed (Maurer 2012; Pignone et al. 2002; Sharp and Lipsky 2002). Because the intended use of screening is not to make diagnosis, making a diagnosis of depression based solely on the scores on a depression screening instrument is an example of misuse. Such a misuse may result in over- or underdiagnosis of the disorder.

Although not explicitly using the term consequences, the International Society for Pharmacoeconomics and Outcomes Research Patient Reported Outcomes Harmonization Group alluded to the issue of consequences in the Ad Hoc Task Force Report on the incorporation of patient perspective into drug development and communication (Acquadro et al. 2003). The report states that "decisions about the incorporation of a PRO strategy into a clinical trial should be made with the research design and intended claim in mind" (p. 527). Questions such as the claim that one is hoping to achieve and the psychometric instruments employed to address the claim need to be taken into consideration. Our findings that consequences were rarely reported suggest that more communication is needed to promote the inclusion of consequences in validation practice.

If inferences and decisions made are not based on scores from instruments with strong psychometric properties, it may lead researchers and medical practitioners to make incorrect decisions. It may also negatively influence the medical diagnoses, treatment interventions, and even the approval of drugs in the market, which in turn may hurt the quality of life of our patients. Just because the authors of a validity study claim that they have validated an instrument and have concluded that the instrument is "valid" does not guarantee that the evidence is adequate to support the inferences made from the scores. Readers and practitioners should always have a critical mind.

The formation of the PRO Content Validity Good Research Task Force (Patrick et al. 2011a, b) to develop good research practices in content validation is encouraging. Perhaps the formation of task forces and making available agreed upon and endorsed best practices and reporting guidelines on other sources of validity evidence (such as response processes and consequences) maybe promising approaches to improving the practice of psychometric validation in health care.

Acknowledgement To obtain a list of the articles included in this study, please contact the corresponding authors.

References

Acquadro, C., Berzon, R., Dubois, D., Leidy, N. K., Marquis, P., Revicki, D., et al. (2003). Incorporating patient's perspective into drug development and communication: An ad hoc task force report of the patient-reported outcomes (PRO) harmonization group meeting at the Food and Drug Administration, February 16, 2001. *Value in Health, 6*, 522–531.

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.

Auewarakul, C., Downing, S. M., Jaturatamrong, U., & Praditsuwan, R. (2005). Sources of validity evidence for an internal medicine student evaluation system: An evaluative study of assessment methods. *Medical Education, 39*, 276–283.

Chan, A. W., & Altman, D. G. (2005). Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet, 365*, 1159–1162.

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.

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.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement, 63*, 65–74.

Hojat, M. (2007). *Empathy in patient care: Antecedents, development, measurement, and outcomes*. New York: Springer.

Hojat, M., Mangione, S., Nasca, T. J., Cohen, M. J. M., Gonnella, J. S., Erdmann, J. B., et al. (2001). The Jefferson scale of empathy: Development and preliminary psychometric data. *Educational and Psychological Measurement, 61*, 349–365.

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.

Kopec, J. A., & Willison, K. D. (2003). A comparative review of four preference-weighted measures of health-related quality of life. *Journal of Clinical Epidemiology, 56*, 317–325.

Maurer, D. M. (2012). Screening for depression. *American Family Physician, 85*, 139–144.

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

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

Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M., & Altman, D. G. (2007). Epidemiology and reporting characteristics of systematic reviews. *PLoS Medicine, 4*, e78.

Patrick, D. L., Burke, L., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., et al. (2011a). Content validity – establishing and reporting the evidence in newly-developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part I – Eliciting concepts for a new PRO instrument. *Value in Health, 14*, 967–977.

Patrick, D. L., Burke, L., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., et al. (2011b). Content validity – establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part 2 – assessing respondent understanding. *Value in Health, 14*, 978–988.

Pignone, M. P., Gaynes, B. N., & Rushton, J. L. (2002). Screening for depression in adults: A summary of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine, 136*, 765–776.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 3*, 385–401.

Richardson, C. G., & Zumbo, B. D. (2000). A statistical examination of the Health Utility Index-Mark III as a summary measure of health. *Social Indicators Research, 51*, 171–191.

Santanello, N. C., Baker, D., & Cappelleri, J. C. (2002). Regulatory issues for health-related quality of life – PhRMA Health Outcomes Committee Workshop, 1999. *Value in Health, 5*, 14–25.

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.

Sharp, L. K., & Lipsky, M. S. (2002). Screening for depression across the lifespan: A review of measures for use in primary care settings. *American Family Physician, 66*, 1001–1008.

Valderas, J. M., Ferrer, J., Mendívil, M., et al. (2008). Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. *Value in Health, 11*, 700–708.

Zillich, A. J., Doucette, W. R., & Carter, B. L. (2005). Development and initial validation of an instrument to measure physician–pharmacist collaboration from the physician perspective. *Value in Health, 8*, 59–66.

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/Boston: 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: Information Age Publishing.

Chapter 16
Validation Practices of the Objective Structured Clinical Examination (OSCE)

Tavinder K. Ark, Neelam Ark, and Bruno D. Zumbo

Introduction

Objective structured clinical examinations (OSCEs) have been used for over a decade in the assessment of medical students and residents. The OSCE is an assessment designed to evaluate the clinical and interpersonal skills of examinees (Harden et al. 1975) because these skills are not measurable through typical written examinations. An OSCE consists of several different short clinical scenarios in which an examine must interact with a standardized patient (SP) exhibiting a chief complaint (such as chest pain).

An OSCE can be a high stakes examination (e.g., licensure exam) or simply used to provide examinees with feedback. There has been some confusion in the literature regarding the measurement validity of the use and interpretation of OSCE scores. Many articles have simply stated that the validity of the OSCE has been well established, referencing less than a handful of articles. Upon closer examination, these referenced articles have not necessarily provided what we would consider validity evidence in the use and interpretation or OSCE scores. In fact, these articles go on to cite other articles that have validated the OSCE. This circular reference perpetuates the claim that the OSCE is valid and has been validated. However, in all its decades of use, only a handful of articles have directly attempted to provide validity evidence for the use and interpretation of the OSCE scores.

Part of this perpetuation of the 'validity claim' comes from the misinterpretation of what constitutes validity evidence. Validity theory has undergone many changes in the last two decades, and many of those changes have yet to be incorporated and utilized in the validation of assessment tools in medical education. For instance, researchers still use face validity as a type of validity evidence; however, it no

T.K. Ark • N. Ark • B.D. Zumbo, Ph.D. (✉)
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_16,

longer is considered a type of evidence in contemporary validity theory. Furthermore, researchers use reliability as the whole of validity evidence, where they claim the OSCE is valid based upon reliability evidence. The researchers of these studies typically rely upon only one source of validity evidence (e.g., discriminant validity); however, a complex assessment such as an OSCE, with its varied uses in different populations, necessitates more than one source of validity evidence and a more elaborated validity argument.

Although it is not often discussed in the field, medical educators should care about validity for many reasons. First and foremost validity can be used to build a legally defensible position for high stakes examinations. There are many public cases where examinees taking high stakes tests have questioned the validity of their test scores. For instance, there have been many cultural and minority discrimination lawsuits in testing. In medical education, many exams, such as the OSCE, are used to pass and fail students and more importantly license physicians to practice medicine. Having validity evidence already established will help build a defensible position for the use and interpretation of test scores for high stakes examinations in medical education.

There were two inter-related purposes for this chapter. First, the current validity evidence and validation practices used for the OSCE need to be re-examined within a contemporary framework of validity. Second, evidence regarding the validity of the inferences from and uses of OSCE scores needs to be re-established in a systematic way using the most current validity standards. With these two purposes in mind, this chapter will compare and contrast the extent to which the reported validity in these papers conforms to modern day validity theory based on the *Standards* for educational and psychological testing (The *Standards*) (APA et al. 1999). Only research articles that clearly position themselves as providing validity evidence will be summarized and investigated in this chapter. This comparison will provide valuable insight and direction for future validation work on the OSCE and the thinking of medical educators on validity in general.

Objective Structured Clinical Examinations

The OSCE is an assessment tool designed to evaluate the clinical and interpersonal skill of medical students, residents and soon to be practicing physicians (Harden et al. 1975). An OSCE is a short clinical scenario in which an examinee interacts with a SP as if they are a real patient. An examinee rotates through a series of OSCE stations (up to 20 cases) and they must attend each station. Each station consists of a medical case where the examinee must deal with a chief complaint such as chest pain. Furthermore, in each case, the examinee has 10 to 15-min to interview and counsel the SP regarding a medical condition or complaint. The OSCE cases are designed to assess the examinee's ability to (a) communicate effectively with the patient (communication skills), (b) gather an accurate history from the patient

(history gathering skills), and (c) perform a focused physical examination on the patient (physical examination skills) (Lawson 2006).

OSCEs are widely used by medical schools and certification organizations not only to track the progress of students and residents, but also to make pass or fail decisions. Therefore, it is not only important that OSCEs are reliable, but that the use and interpretation of test scores are valid.

Sources of Validity Evidence

According to Messick (1989) and the *Standards* (APA et al. 1999), validity refers to the degree to which evidence and theory supports the interpretation of test scores as outlined by the proposed use of a test about individuals from a given sample and context. Validation refers to the ongoing process of accumulating evidence to support the proposed interpretation and use of test scores. A measure, per se, is not validated, but rather the proposed interpretation and use of test scores one makes from a measure are validated. The inferences one makes from a test scores is bound by the context, the population and the use of the test scores.

The proposed interpretation of test scores is not the same as the intended use of test scores and both need to be explicitly articulated and stated before the validation process can begin. The interpretation of test scores refers to the construct the test is intended to measure. The intended use of the test scores implies different interpretation of the test scores. This is why the interpretation of the test scores is entailed by the proposed use of the scores, and why both need to be explicitly stated. In the validation process the degree and accumulation of evidence needs to support the interpretation of the tests scores as well as the intended use of test scores. Validation also involves examining the ways in which possible distortions to the interpretation and use of test scores may arise. This may involve providing further evidence for the proposed interpretation and intended use of the test scores.

The dominant current conceptualization is that construct validity is the whole of validity and that validity is a unitary concept that utilizes multiple sources of evidence to support the proposed interpretation and intended use of test scores. Construct validity refers to the degree to which the observed scores measures the underlying theoretical constructs those observations are meant to capture. Two major threats to construct validity that must be considered are (a) construct underrepresentation and (b) construct irrelevant variance. Construct underrepresentation refers to instances when a test does not capture all or important aspects of the construct being measured (Sireci 2009). It also includes narrowed meaning of test scores. For instances, the OSCE may be missing some vital skills or competencies required for practicing medicine such as cross-cultural competency (this refers to the ability of individuals to interact and communicate with individuals of different cultural backgrounds). Construct irrelevant variance refers to instances when the tests scores are affected by other extraneous variables other than the construct being measured. This also includes instances when the test scores may be influenced

systematically by other components that are not part of the construct (Sireci 2009). For instance, the OSCE could be capturing additional constructs such as language comprehension, which is not the primary focus of the OSCE. Therefore it is important to ensure the evidence of the test content adequately represents, instantiates, and is relevant to the construct being measured.

The *Standards* (APA et al. 1999) outline five distinct sources of validity evidence to support the proposed interpretation and use of their assessment tools. The five sources of evidence are: content, response processes, internal structure, relationships to other variables, and consequences. Each one of these sources of validity evidence will be defined and discussed in the context of medical education, specifically using the OSCE.

Evidence based on content is similar to traditional content validity (Sireci 2009). Evidence based on content involves the use of subject matter experts to review, rate and deliberate the inclusion of test items. Part of this process involves the various experts discussing how the content domain should be represented and the relevance of the items in capturing the content domain and the test specifications. More complex analyses can be used such as evaluating the link between the content in the curriculum to the content of the test (APA et al. 1999).

Evidence based on response processes refers to the extent to which processes (cognitive or behavioral) are consistent with the intended interpretation of test scores. In other words, it is the degree of fit between the construct and the activity the examinees engage in (APA et al. 1999; Messick 1995). For instances, using think-a-loud protocols may provide information regarding the reasoning processes used by examinees when solving an item on a test. If the examinees reasoning process matches the way in which developers expected the item to be answered, then this provides evidence for response processes.

Evidence based on internal structure refers to the degree to which test items and the sub-components conform to the construct on which test scores interpretations are based (APA et al. 1999; Sireci 2009). This involves using statistical analyses of item, sub-scores, and sub-scales to investigate the dimensionality of the latent variable that is being measured. There are a number of statistical procedures that can capture such information such as Confirmatory and Exploratory Factor Analysis (CFA and EFA), multidimensional scaling, Classical Test Theory (CTT) and Generalizability Theory (GT). It is important to realize that CTT is not the only statistical technique available to investigate the internal structure of items on an assessment.

Evidence based on relationships to other variables examines the degree to which relationships between the test scores and other variables are consistent with the construct underlying the proposed test interpretations (APA et al. 1999). The relationship to other variables can be broken down further into convergent, discriminant, test-criterion and validity generalization depending on what variables are related to the test score. Convergent evidence refers to the relationship between test scores and variables intended to measure similar constructs (APA et al. 1999). Discriminant validity not only refers to relationships between test scores and variables intended to measure different constructs, it also refers to evidence when the

test scores differ significantly across groups as it is expected or hypothesized (i.e. level of expertise, or experimental condition versus the control) (APA et al. 1999). Evidence based on test-criterion refers to the relationship between the test-score and a relevant criterion measure – this criterion measure is of primary interest. This may include predictive studies if the test score is being used to predict a criterion score. Validity generalization evidence refers to instances were the test score is being used to predict the same or a similar criterion in a different context (APA et al. 1999).

Finally, the evidence based on the consequences of testing refers to the evaluation of intended and unintended interpretation and use of test scores (APA et al. 1999; Hubley and Zumbo 2011). The evidence based on consequences is not the same as test misuse. Misuse refers to consequences of unsound interpretation; procedural errors and illegitimate uses of test scores (Messick 1998).

The *Standards* (APA et al. 1999) provide not only definitions of the five sources of validity evidence, but guidance as to what type of validity evidence is needed to validate the interpretation of test scores. Outlining and defining the sources of validity evidence will help in understanding the evidence and perspective that authors adopt and report when analyzing the validity studies on the OSCE.

Research Questions

The aim of this study is to examine the validation practices and validity evidence that has been reported in regards to the inferences from and use of OSCE test scores. Two inter-related research questions guided our study.

1. What is being reported as validity or validation evidence on the OSCE test scores?
2. To what extent do these studies reflect and conform to contemporary validity theory perspectives according to the *Standards* (APA et al. 1999) and Messick (1989)?

The answer to these questions will provide valuable guidance and inform future validity studies of the OSCE.

Methods

Search Strategy and Study Selection

The PubMed database from January 1966 to June 2013 was used to search for articles examining the validity or validation of the OSCE. Pubmed was used because it is one of the most common and widely used free online databases

available. It accesses articles from the MEDLINE database and is operated by the United States National Library of Medicine (NLM) at the National Institute of Health (NIH). It is one of the primary databases used by health care professionals, medical educators and researchers in the medical field.

Peer reviewed studies were included in the data analysis if they were (a) primarily interested in investigating the validity or validation of the OSCE, and (b) were explicitly presenting their findings as validity. This was achieved by searching for articles that had the term 'validity' or 'validation' in the title of the article, and either the term 'OSCE' or 'objective structured clinical examination' in the title or abstract. Only articles that met this criterion were included in this study. Furthermore, only articles that examined the validity or validation of the OSCE in the health care fields (such as medicine, dentistry, or nursing) were included in the data analysis. All of the articles were scrutinized to ensure the OSCE assessment was from the health care field, and that the primary interest was validity or validation of the OSCE.

Data Collection and Analysis

A similar strategy to Cizek et al. (2008, 2010) was used to extract, code and characterize validity evidence in the selected articles. Some modifications were made to Cizek et al.'s (2008, 2010) coding scheme so that additional information relevant to validity and validation of the OSCE scores could be coded. For instance, Cizek et al. (2008, 2010) coded whether the internal structure was reported as validity evidence, reliability evidence or as reliability evidence bearing on validity. Additional categories were added to Cizek et al.'s (2008, 2010) such as whether authors of the selected article reported internal structure as validity when it is reliability, or reliability evidence as validity.

All variables were coded dichotomously as being present or not. The following validity evidence was coded:

(a) Validity Perspective. Articles were reviewed for whether the authors of the selected articles provided a unitary validity perspective, cited contemporary validity references (e.g., the AERA *Standards* or articles by Messick, Cronbach and Meehl, Zumbo, or Kane), and referred to validity as a characteristic of the test or as the characteristics of the inferences/scores, or both. If no clear validity perspective was provided or it could not be ascertained from the article, the article was coded as 'unclear or not present' in providing a validity perspective.
(b) Sources of Validity evidence. Each article was analyzed and coded for the sources of validity evidence based on the *Standards* (APA et al. 1999). Evidence was coded as any of the following: test content, response processes, internal structure, relationship to other variables (convergent, discriminant, test-criterion, and validity generalization) and the consequences of testing. For ease of coding, the relationship to other variables category was subdivided

into its traditional forms of criterion-related evidence for validity such as predictive, concurrent and divergent validity. Face and construct validity were also included to help code what was reported by the authors of the selected articles. Internal structure was coded in two ways. First, internal structure was coded in terms of how it was presented, which was as follows: validity evidence only, reliability evidence only, as reliability and validity evidence, as validity when it should be reliability, or as reliability when it should be validity. Second, internal structure was coded as being present if some sort of statistical analysis was provided to determine if the items on the test were measuring the appropriate construct.

(c) Types of reliability presented. Each article was analyzed and coded for the type of evidence the authors provided as reliability. Reliability evidence was coded as internal consistency, parallel forms, test-retest, and as intra/inter-rater or – station reliability.

(d) Validity references. Articles were examined and coded for whether or not a contemporary validity article was referenced. If an author outside of the contemporary view of validity was cited and used to provide a validity framework, it was documented and reported separately.

Two rounds of coding occurred. In the first round, validity perspective, sources of validity evidence, types of reliability presented and validity reference was reported as the author presented it in the article. In the second round of coding, the sources of validity evidence were reanalyzed and re-coded using the *Standards* (APA et al. 1999). Therefore, the way in which the authors of the selected article reported validity may not coincide with how the *Standards* (APA et al. 1999) would classify or report the same validity evidence.

The purpose, use and interpretation of test scores can vary across medical programs and examinees. As a result, additional information regarding the inferences, skills measured, medical domains assessed and examinees was analyzed and coded. This information will provide valuable insight into understanding under what context, participants, and medical domains the validity results of these papers can be generalized to. In addition, this information provides a context in understanding how and why the authors of the selected papers validated and reported validity the way in which they did. The coding was qualitatively driven based on the information provided in the article. The additional information analyzed and coded from the selected articles were as follows:

(a) Use or inferences being made from the OSCE. The purpose of the OSCE or inferences made from OSCE test scores were coded. This provides important information regarding whether the authors of the selected article validated their reported use and interpretation of the OSCE test scores.

(b) Skills measured by the OSCE. Articles either listed a specific skill, such as performing a particular physical examination procedure, while others listed more generic skills, such as clinical competence. Articles were coded as either presenting a generic or a set of specific skills that the OSCE measured.

(c) Medical domains assessed by the OSCE. Each article was coded for what medical domains the cases or stations were measuring. Some OSCEs used a wide variety of cases including a wide variety of cases, such as general internal medicine, psychiatric and surgical cases. Other OSCEs consisted of entirely one medical domain, such as surgical cases. The selected articles were coded based on the information the authors provided regarding the medical domains assessed by the OSCE cases.
(d) Participants. Each article was coded for the types of participants (e.g., undergraduate, resident or practicing physician) that were used in the validation of the OSCE test scores. This provides information regarding what populations the validity results can be generalized to.

The reliability of the coding was examined for all selected articles. The agreement between the first author and an independent rater was very high, with judgments for sources of validity reaching 93 % for exact agreement on the selected articles – the agreement was 91 % for the first round examining what authors reported and 94 % in the second round of coding using the *Standards* (APA et al. 1999). Any discrepancies in coding between the first author and an independent rater were discussed until a consensus were reached on how the article should be coded.

Results

Study Selection and Search Strategy

A total of 34 articles were found that contained 'validity' in the title and the term 'OSCE' or 'objective structured clinical examination' in the title or abstract. Of these 34 articles, a total of 16 articles were excluded from the analysis. That is, three articles were excluded because of access issues – the original articles could not be obtained and the abstracts lacked the detail to analyze and accurately code the validity evidence. An additional nine articles were excluded because the OSCE was being used to validate another assessment tool via correlations or were completely unrelated to the validity of the OSCE. Three more articles were excluded because the OSCE was used to assess areas outside the medical domain such as audiology, physical therapy or midwifery. In addition, the Hodges (2003) paper was excluded because it was a conceptual paper regarding the validity of the OSCE. Therefore, a total of 18 articles were included from this initial search.

In another search, a total of nine articles were found that contained the term 'validation' in the title and the term 'OSCE' or 'objective structured clinical examination' in the title or abstract. Of these articles, only four were used in the analysis because the remaining five articles did not examine explicitly examine the validation of the OSCE. That is, three articles were interested in using the OSCE as a gold standard to validate another assessment tool, one article could not be

Table 16.1 Summary of the domains assessed by the OSCE

Medical domains assess	n	% of articles
Endocrine	1	4.5
Variety of cases (internal medicine, clinical rotation)	14	63.6
Psychiatry	2	9.1
Musculoskeletal	1	4.5
Surgery	1	4.5
Dentistry	2	9.1
Ophthalmoscopy	1	4.5

obtained due to access issues, and one article examined the validation of a simulation OSCE.

Therefore, the first search resulted in 18 and the second search an additional 4 for a total of 22 articles that were included, analyzed and coded for the validity evidence regarding the OSCE.

Data Collection and Analysis

Each study was coded for the perspective of validity presented, sources of validity evidence, types of reliability evidence and cited validity references. Table 16.1 summarizes the medical domains that are being assessed by the OSCE. The articles mostly examined OSCE scores from undergraduate medical students (54.5 %), followed by residents (36.4 %) and then practicing physicians (13.6 %).[1] This suggests most OSCE are designed for undergraduate medical students and residents, but not practicing physicians. The OSCEs represented a breadth of medial domains including very specialized areas (e.g., surgery, psychiatry and musculoskeletal systems). However, for the most part, the OSCEs in these selected articles consisted of a wide variety of cases (63.6 %). This suggests not one area of medicine has a plethora of studies on the validity of the OSCE and it adds to the breadth of validity analysis. Most of the selected articles reported using the OSCE to assess or grade medical students/residents (31.8 %), certify physicians (22.7 %), aid in the research and development of the OSCE cases (18.2 %) and assess, evaluate, measure competency and skills of the participants who take them (18.2 %). Only two articles did not specify the inferences that are being made from the OSCE (9.1 %). With respect to the skills assessed, 63.6 % of the selected articles listed the specific skills the OSCE captured, such as performing a specific physical examination skill, while 36.4 % of the articles generically stated the OSCE measured clinical competence, skills or performance.

[1] Numbers do not add up to 100 % because some articles used both resident and undergraduate medical students.

Table 16.2 Summary of validity perspective markers

Validity perspective markers	n	% of articles
Unitary perspective stated	2	9.1
Standards (APA et al. 1999) or Messick (1989) cited	2	9.1
Conception of validity		
As a characteristic of the test	3	13.6
As a characteristic of the test score or inferences	5	22.7
Both	4	18.2
Unclear/neither	9	40.9

To answer the first research question, each article was analyzed and coded based on what authors reported as validity/validation evidence and the validity perspective they provided on the OSCE. To assess the validity perspective provided by authors, three markers were used. The first marker examined whether the authors of the article presented a unitary perspective of validity over the different 'types' of validity (Sireci 2009; Cizek et al. 2008, 2010). The second marker examined whether the authors of the article referenced the *Standards* (APA et al. 1999) or papers published by Messick on validity. The last marker examined whether the authors of the articles presented validity evidence as a characteristic of the test, inferences, both or was unclear. Some authors oscillated between validity being a property of the test and a property of the score inferences. In these particular instances, the article was classified as the authors reporting validity as a characteristic of the test and inferences.

Table 16.2 summarizes the validity perspective reported by the authors of the selected articles. In general, the modern view of validity is not discussed or reported by the authors of the selected articles in validating the OSCE scores. Only two articles (Auewarakul et al. 2005; Varkey et al. 2008) explicitly stated and were guided by the unitary view of validity. In particular, the Auewarakul et al. (2005) article cited Messick (1989) and the *Standards*. Most of the articles used language to imply validation involved providing "a type" or "types of" validity evidence and presented their results as such.

With respect to the conception of validity, many of the authors of the selected articles referred to validity as a characteristic of both the OSCE and the OSCE scores or inferences. Furthermore, validity was more commonly referred to as a characteristic of the OSCE scores and inferences, than a characteristic of the OSCE scores or inferences in the selected articles. Most of the authors did not give a clear perspective regarding their conception of validity. This lack of specification and confusion between whether validity is a characteristic of the OSCE or the OSCE scores suggests that there may be confusion surrounding the conception of validity.

To examine the second research question, the validity evidence in the selected articles was coded using the five sources of validity evidence listed in the *Standards* (APA et al. 1999). These sources included evidence based on test content, response processes, internal structure, relationships to other variables and consequences of testing. Many of the authors of the selected articles reported

Table 16.3 Summary of sources of validity evidence reported

Types of validity evidence based on author reporting	n	% of articles
Face	6	27.3
Content	9	40.9
Construct	13	59.1
Response processes	2	9.1
Relationships to other variables	1	4.5
Consequences	2	9.1
Criterion-related		
Predictive	5	22.7
Concurrent	9	40.9
Internal structure presented		
As reliability (only)	10	45.5
As reliability and validity	6	27.3
As validity (only)	0	0.0
As validity, when it is reliability	8	36.4
As reliability when it is validity	1	4.5

validity evidence that existed in the previous views of validity theory such as face and construct validity, which no longer exist in the *Standards* (APA et al. 1999) or modern day thinking of validity. However, for the sake of accuracy, these categories were included to accurately code what authors were reporting as validity evidence on the OSCE. Furthermore, certain categories in the *Standards* (APA et al. 1999) were further subdivided to include previous views of validity. The evidence based on the relationships to other variables was subdivided into criterion-related predictive and concurrent evidence. Therefore, the final categories used to analyze the articles based on what was reported by the authors were as follows: (a) face validity, (b) content validity, (c) response processes, (d) internal structure, (e) criterion-related predictive validity, (f) criterion-related concurrent validity, (g) relationship to other variables, (h) consequences of testing, and (i) construct validity.

The sources of validity evidence reported by the authors in the articles are summarized in Table 16.3. The most frequently reported source of validity evidence on the OSCE was construct, followed by content and criterion-related concurrent validity. The remaining sources of validity evidence were reported less frequently, which included face validity, criterion-related predictive validity, response processes, and consequences. The one study that reported relationship to other variables used the *Standards* (1999) in guiding the validation of the OSCE; however this article did not specify the relationship of the variable that the OSCE was being correlated to. Reliability evidence was mostly reported as reliability evidence. There were eight articles where reliability was reported as validity and one article where validity was reported as reliability. In the cases where reliability was reported as validity, the evidence based on the internal structure was confused as validity. All of the internal structure evidence provided by the authors of the selected articles provided reliability values (e.g., inter-rater, internal consistency)

Table 16.4 Summary of sources of validity evidence using the Standards (APA et al. 1999)

Types of validity evidence	n	% of articles
Content	5	22.7
Response processes	2	9.1
Consequences	1	4.5
Relationship to other variables		
Convergent	10	45.5
Discriminant	12	54.5
Test-criterion	3	13.6
Validity generalization	5	22.7
Internal structure	9	40.9

derived from classical test theory or G-theory except for one article. One article used confirmatory factor analysis to provide the internal structure of the OSCE.

With respect to the actual validity evidence provided, 17 of the 22 articles reported the OSCE to be valid based on whatever sources of validity evidence the authors provided. Three of these articles stated that more validity evidence was still required in the validation of the OSCE. One article reported negative findings and five had mixed or neutral feelings regarding the validity of the OSCE. Of the studies that reported positive, negative and mixed/neutral findings, 6 of the articles explicitly stated further work on the validation of the OSCE is needed before the OSCE scores can be used to make inferences regarding the examinees.

The data summarized in Tables 16.2 and 16.3 suggest that there may be a mismatch between the reported evidence and the *Standards* (APA et al. 1999). Therefore, in order to further explore the second research question, each of the selected articles sources of validity evidence were reanalyzed and recoded based on the definitions of validity evidence from the *Standards* (APA et al. 1999). The validity evidence based on the *Standards* (APA et al. 1999) includes test content, response processes, internal structure, relationships to other variables and consequences of testing. The category, 'relationships to other variables' was subdivided into the categories provided in the standards, which are: (a) convergent validity, (b) discriminant validity, (c) test-criterion validity, and (d) validity generalization. Under the current *Standards* (APA et al. 1999) anything classed as construct validity that examined the group differences (such as the difference in OSCE scores between 1st year residents to 4th residents) was classed as discriminant validity. Criterion-related predictive and concurrent validity were reclassified into either convergent, test-criterion, validity generalization categories based on what was being correlated to the OSCE scores. Face validity was removed from this analysis because it is not in the current *Standards* (APA et al. 1999).

Table 16.4 summarizes the sources of validity based on the *Standards* (APA et al. 1999). A significant source of validity evidence regarding the OSCE was discriminant, convergent and internal structure validity. The remaining sources of validity was reported far less and included content, validity generalization, test-criterion, response processes and consequences.

Figure 16.1 presents the differences between what authors' reported as validity evidence to how the same evidence could be classified using the *Standards* (APA

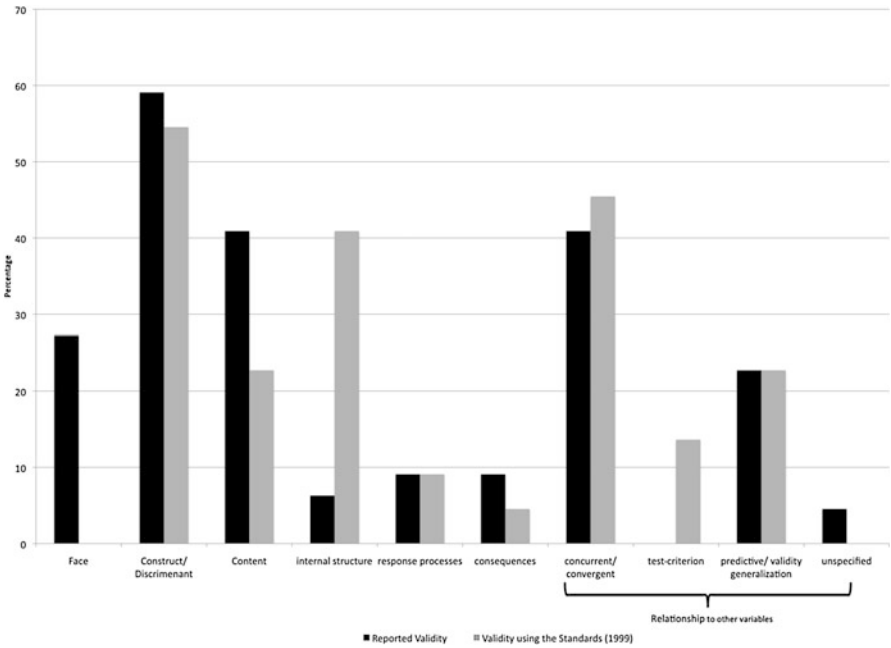

Fig. 16.1 The differences between what authors' reported as validity evidence to how the same evidence would be classified using the Standards (APA et al. 1999)

et al. 1999). Two notable differences were found. First, there was a significant decrease in the articles reporting content validity compared to what the *Standards* (APA et al. 1999) and Sireci (1998) would describe as content validity. More than half the authors reported content validity; however, after close examination only 22.7 % were providing content validity as described by the *Standards*. The second notable difference is the increase in the number of articles reporting convergent validity based on the *Standards* (APA et al. 1999). Just less than half the articles reported convergent validity evidence based on the *Standards* (APA et al. 1999), whereas 40.9 % of the authors were reporting some form of criterion-related evidence based on Table 16.3. That being said, based on the *Standards* (APA et al. 1999), over 90 % of the articles (n = 20) provided some form of evidence based on the relationships to other variables as a source of validity evidence. Likewise, 90 % of the papers (n = 20) reported some form of criterion-related concurrent, predictive or construct validity when adding up the values from Table 16.3. Therefore, the numbers are not so different from what is reported as criterion-related and construct validity evidence by the authors of these papers to what would be classified as relationship to variables and its sub-categories based on the *Standards* (APA et al. 1999). In essence, it is a relabeling of validity evidence to match up with contemporary validity theories.

Table 16.5 summarizes the types of reliability evidence provided by authors of the selected articles. Ninety-six percent of the articles presented some form of

Table 16.5 Summary of reliability evidence reported

Types of reliability evidence	N	% of articles
Inter-rater/station	9	40.9
Test-retest	0	0.0
Internal consistency	9	40.9
Parallel forms	0	0.0
G-theory	6	27.3
Unspecified	4	18.2
Total number of articles reporting some form of reliability	21	95.5

reliability data regarding the OSCE. Most articles reported inter-rater or –station reliability and internal consistency of items. In many of the cases a coefficient alpha was reported for internal consistency, but it was not clear if the internal consistency of items was done across all items on a checklist, cases or sub-domains, and then averaged across to provide one value for internal consistency. No test-retest or parallel forms reliability was conducted. Six of the 22 selected articles conducted G-theory. However, for at least three of the articles it was unclear how the G-coefficients were derived – for example, information regarding which factors were included as fixed or random were not provided. Only one article provided variance components for each of the sources of variance in the analysis. Surprisingly, four of the articles did not specify what type of reliability indexes they conducted or quoted previous studies that had shown the OSCE to have reliability.

Lastly, seven of the 22 articles (31.8 %) cited validity articles, which included Messick's (1989) seminal chapter on validity, the *Standards* (APA et al. 1999), Kane's (1982) paper on the validity of licensure exams, and Anastasi's (1982) book on psychological testing. Another article that was cited with respect to validity was Downing (2003). Downing's (2003) article does an excellent job of summarizing Cronbach and Meehl's (1955), Messick's (1989) and the *Standards* (APA et al. 1999) in the context of medical education. Moreover, Downing's (2003) article also used the OSCE as an example to frame how validity studies can be conducted in medical education.

Discussion

Based on the sources of validity evidence provided, many of the articles claimed the OSCE to be a valid and reliable tool. Fifty-five percent of the articles provided some sort of definitions of validity and the various "types" that they reported in their paper for validating the OSCE. By doing this, the authors of these selected paper provided the readers with a framework of validity theory they were working from. However, the remaining 45 % of articles did not provide a definition of validity and for the "types" of validity evidence they provided in their articles. This often can cause confusion in how the authors classified the sources of validity evidence they

provided on the OSCE because the reader did not always know what validity framework the authors were using.

The discussion section will be divided into three parts. The first section is a critique of validity evidence provided by the authors. This section will also examine how the confusion in the definitions of validity evidence may lead to the inaccurate representation of validity and validation of the OSCE. The second section will discuss the issues surrounding the reporting of reliability. In particular, this section will discuss how reliability is not enough to prove the internal structure of the OSCE and that reliability should not be confused 'as' validity evidence, but considered a piece of evidence in the argument for validating the OSCE. The final section will provide future direction for validity studies on the OSCE in medical education.

Critique of the Reported Validity Evidence

Content Validity

Just less than half the articles reported content validity. Using the definitions and methods outlined by the *Standards* (APA et al. 1999) and Sireci's (1998), half of these articles did not provide adequate information regarding how they evaluated the content in their OSCEs. That is, after closer examinations of what authors reported as validity evidence in their articles, more than half of the articles were actually not accurately reporting content validity. No description was provided in terms of who the experts were, how a consensus was achieved and what types of judgments or feedback was given by the experts on the content of the OSCE.

In some articles, experts and OSCE examinees were asked if they felt the OSCE represented and captured the courses and the objectives of the curriculum using surveys or oral feedback. This was considered evidence for content validity. Other examples provided by authors as content validity, when it is not based on the *Standards* (APA et al. 1999) and Sireci's (1998), are if experts or participants felt (a) the clinical cases in the OSCE were believable, (b) seen frequent in practice, (c) the duration of the case is adequate, (d) the SP portrayal of the case was accurate, and (e) the cases in an OSCE assessed the skills required to practice medicine. A lot of these judgments were based on expert or participants opinions, but very little analysis was provided as to how the content of the OSCE measured the skills or competency they expected the OSCE to capture. Much of the content evidence provided by authors was evidence for quality control than evidence for content. A huge caveat is that most of disqualified articles simply lacked reporting the methods they adopted in assessing the content validity of the OSCE. That is, the OSCE in the selected articles may have undergone more rigorous analysis of content validity; however, based on the evidence the authors reported in the articles, one cannot accurately discern if this was the case.

Based on the *Standards* (APA et al. 1999), test content refers to the themes, wording, format, questions and tasks as well as scoring. According to Sireci (1998), content validity refers to the extent to which a test measures the content domain it purposes to measure. To provide evidence for the test content, developers must show the items and tasks on the test are representative of the targeted content domain. Combining this definition with the *Standards* (APA et al. 1999), content validity can be seen as examining the relationship between the content and the construct that is being measured by the OSCE. This can be achieved by defining the medical domains or construct that the OSCE is measuring, followed by assessing the extent to which the OSCE cases match the definitions and the items on the checklist that are relevant to assess the medical domains. Experts can be used; however, information needs to be provided in terms of (a) who the experts are, (b) how they were selected, and (c) what type of information the experts provide in the selection of items and construction of the OSCE content.

Of the articles that provided evidence for test content, these articles utilized experts to create a grid listing the medical content domains against the skills needed for each domain. This grid was then aligned with the learning outcomes of the curriculum. In order to ensure the OSCE captured the relevant content and constructs, the content domains (e.g., hearth attack) are plotted along one axis, with skills plotted against the next (e.g., patient education). Each OSCE station is mapped back onto this grid to ensure that the OSCE covers every domain and skills the participants are expected to know. In other cases, evidence for the test content was established using experts to compare the content and skills assessed in an OSCE with the curriculum objectives in a very similar manner. Although neither of these methods are perfect, it was the closest examples of content validity based on what the *Standards* (APA et al. 1999) and Sireci (1998) would classify as evidence towards test content.

Face Validity

Face validity is not included as a source of validity evidence by contemporary validity theorists and the *Standards* (APA et al. 1999). Six articles explicitly reported face validity, with 4 of the articles published after 2005, while 2 were published before 1991. Considering that the *Standards* were published in 1999, it is still surprising to observe researchers report face validity as a source of validity evidence. Face validity examines if the assessment tool 'looks' or 'feels' as if it assess what it is supposed to measure (Anastasi 1986). For instance, a math test can be said to have face validity if the test looks like it has math problems on it. Likewise, an OSCE is said to have face validity because it looks like it is measuring physician skills and that it resembles a real life patient-physician encounter.

The use of face validity brings up a debate regarding fidelity. Fidelity and face validity are not to be confused with one another. Fidelity refers to 'exactness.' However, in simulation studies, fidelity refers to the degree to which a model or

simulation reproduces the behavior of a real world object or person, or condition (Weller et al. 2003). Therefore, fidelity is a measure of realism or similarity in terms of mimicking a real life situation. This is very different from face validity, which examines whether a measure "feels like" or "looks like" it is measuring what it is suppose to measure. Fidelity examines exactness or similarity between two situations in hopes to evoke similar behaviors as a real life situation.

Part of the challenge to both face and content validity is that experts or those who are invested (e.g., experts who work at the medical school or OSCE participants) in the development of the OSCE are selected. As a result, they can indirectly provide a biased opinion in terms of the OSCE content being representative of the curriculum objectives or the skills they want the OSCE to assess. These individuals are invested in the OSCE. As a result, experts should be included from a broad range of backgrounds, including ones that are not involved in the development or writing of OSCE cases, to provide unbiased and accurate feedback regarding the test content.

Criterion-Related Convergent Validity

Articles reporting criterion-related convergent validity correlated the OSCE scores to various assessment tools. In general, the articles reported correlations ranging from as low as 0.19 to as high as 0.68. The OSCE scores were correlated with (a) various types of clinical skills exams, such as the mini-CEX, bedside examinations, and other types of OSCEs, (b) certification scores or final-year grades, (c) other test scores, such as knowledge or problem solving scores, (d) global ratings to sub-domains on the checklist of the OSCE, and (e) patient ratings. The wide variety of assessment tools that the OSCE scores were correlated to suggests a lack of a gold standard or criterion in the field. It also illustrates the lack of consensus of what the gold standard or criterion should be to compare or validate the OSCE scores. This may be attributed to the fact that the OSCE is used for a wide variety of reasons, various purposes and can measure different constructs. As a result, a gold standard cannot be used to validate the OSCE since its purpose, use and the constructs it measures can vary drastically. In addition, the articles did not make an argument as to why they correlated the OSCE scores to a particular measure. In some cases, the construct measured in the OSCE and the correlated assessment tool were not the same, and yet the authors still reported the correlation as convergent or concurrent validity.

The purpose, use and interpretation of OSCE scores and even the reported constructs it measures vary from medical school to medical school. Each medical school and licensure exam boards have created their own silos of OSCE assessment, with different scoring rubrics, constructs and feedback. This poses a challenge in interpreting and understanding the results published as validity studies on the OSCE. This also means each school may have to validate the inferences and uses of their own OSCE and cannot generalize the findings published by other authors unless the OSCEs are similar if not identical.

Construct Validity

The majority of papers (n = 13) that reported construct validity defined it as the degree to which assessments can discriminate between different groups. In these studies the authors provide evidence for construct validity by comparing the differences in OSCE performance by participants in different years or levels of training (i.e., comparing interns to 3rd year residents). The reason authors believe this to be construct validity is because the OSCE is designed to measure clinical competence and skills, and if this is true one would expect those with more advanced training to perform significantly better than those with very little training.

However, this type of thinking fails to take into account construct irrelevance and construct under-representation. The OSCE could be measuring additional constructs, such as cultural sensitivity, that explain for the variance observed in the OSCE scores. For example, an examinee may not be able to effectively communicate with the SP with a different cultural background. As a result, the examinee performs poorly because they could not take the SP's history accurately because of a communication barrier. Also, the OSCE cases may not fully represent or capture all aspects of the construct that is being measured either. Capturing the skills necessary to practice medicine is complex and multifaceted. It is hard to represent a full spectrum of all the skills and competencies needed to practice medicine. Therefore, it is a big assumption to assume the OSCE can assess every single construct, competency or skill needed to practice medicine.

The Problem of Reliability in Understanding Validity

Reliability is defined as the degree to which test scores are free from measurement error (Arnold 1996), In measuring a construct two sources of variance are examined: systematic and unsystematic. Systematic variance represents variability that is due to real differences; however, unsystematic variance is unintended and is unique to the measurement or sample (Hoyt and Melby 1999). For example, systematic variance would be the examinee's communication scores on the OSCE, while SP bias or examinee fatigue would be examples of unsystematic variation. Determining the reliability of a measure is a fundamental way to reflect the amount of error (systematic and random) in a measurement. The reason why researchers are often worried about error is that it leads to the attenuation of the observed score (Hoyt and Melby 1999). This attenuation of the observed score leads to underestimations of the true relation between the construct that is being measured, and the tool designed to capture it.

Every article except for one reported some form of reliability evidence. Most articles provided inter-rater or inter-station reliability and internal consistency. Only six of the articles that reported reliability evidence interpreted this information to provide reliability and validity evidence. Eight of the articles

explicitly confused reliability for validity evidence, instead of stating reliability is a part of the validity argument. Perhaps some of the confusion is the result of the definitions, or lack thereof, of validity and reliability.

The remaining articles indirectly implied that reliability is the most important aspect to justifying the use of the OSCE – if the OSCE could be shown to reliably discriminate between participants, then the authors claimed the OSCE was a reliable tool and should continue to be used. Finally, other statistical analyses than coefficient alpha, such as factor analysis, need to be explored to investigate whether the items are measuring the same underlying variable – i.e., the dimensionality of the items or tasks.

The Future of Validity and Validation Studies Regarding the OSCE

It is noteworthy that there are two articles – Auewarakul et al. (2005) and Varkey et al. (2008) – that used the current *Standards* (APA et al. 1999) to inform their validity studies. Both of these articles provided a unitary perspective of validity and discussed five sources of validity evidence listed in the *Standards* (APA et al. 1999). The study by Auewarakul et al. (2005) sought specific sources of validity evidence to support or refute the proposed interpretation of OSCE scores. The only source of evidence they were unable to provide was consequential evidence because the authors felt the impact of the OSCE on each individual could not be determined. With respect to validity evidence, the authors provided information on the internal structure, responses processes and relationship to other variables.

In contrast, Varkey et al. (2008) provided evidence for each of the five sources of validity evidence including consequences. Specifically, the authors of this article provided content, response processes, internal structure, consequences, and discriminant validity with respect to relationships to other variables. Furthermore this study was conducted explicitly as a pilot study to determine the psychometric and validity of an OSCE used to assess the competencies of problem solving and system based practices. Both the Auewarakul et al. (2005) and Varkey et al. (2008) provided adequate evidence in the first step to validating the OSCE.

Both of these articles were explicit in stating their perspective on validity was a unitary one. In addition, the authors of these articles explicitly stated the purpose of their study was to provide validity evidence regarding the inferences and uses of the OSCE scores, not the OSCE itself. Medical education has much to learn from these two articles in terms of validity and validation, and this is true beyond just validating the OSCE. These two studies are good examples of how valuable it is to incorporate validity theory and the *Standards* (APA et al. 1999). In addition, it also provided a common definition and understanding regarding the various sources

of validity evidence (i.e., discriminant and test content validity). Articles like these will help move the field forward.

One of the limitations of this chapter is that only articles that explicitly presented themselves as validity or validation of the OSCE were analyzed. The purpose for creating such a tight exclusion criterion was to examine articles that were explicitly presenting their results as validity evidence for the OSCE and could not be confused as anything else. In doing this, it may have resulted in the exclusion of articles that examined the validity of the OSCE even though it may have not been the authors main focus or they may not have presented their findings as such.

The larger issues at hand is that many of the articles have differing ways in which the OSCE is being used, what inferences are being made from the OSCE scores, and what constructs the OSCE scores capture. Many papers were trying to validate the OSCE more generically, and to be used in a wide variety of contexts and reasons. The only thing the field seems to agree upon is that the OSCE is supposed to measure clinical competencies and skills. If the specific constructs are listed as to what the OSCE is capturing, the constructs are rarely defined, leaving the reader to assume what certain constructs means such as professionalism. Furthermore, the use and interpretation of the OSCE scores varies across institutions, states and countries. Even the constructs vary in terms of what the OSCE is capturing. This is very important as the inferences are validated, not the OSCE itself. As a result, the OSCE needs to be validated for each way it is being used and cannot be generically validated. Therefore, it is imperative that the inferences that are being made from the OSCE are explicitly stated to help ensure the validation argument matches the inferences and use of the OSCE scores. This also limits the generalizability of OSCE validity evidence. Even though other authors have proved their use and interpretation of the OSCE scores is valid, it does not imply that the OSCE in general is a valid tool.

References[2]

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

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

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

Arnold, M. E. (1996). Influences on and limitations of classical test theory reliability estimates. *Research in Schools, 3*, 61–74.

*Auewarakul, C., Downing, S. M., Jaturatamrong, U., & Praditsuwan, R. (2005). Sources of validity evidence for an internal medicine student evaluation system: An evaluative study of assessment methods. *Medical Education, 39*, 276–283.

[2] *References marked with an asterisk indicates studies included in this review.

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.

*Cohen, R., Reznick, R. K., Taylor, B. R., Provan, J., & Rothman, A. (1990). Reliability and validity of the objective structured clinical examination in assessing surgical residents. *The American Journal of Surgery, 160*, 302–305.

*Cohen, R., Rothman, A. I., Poldre, P., & Ross, J. (1991). Validity and generalizability of global ratings in an objective structured clinical examination. *Academic Medicine, 66*(9), 545–548.

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

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

*Gerrow, J. D., Murphy, H. J., Boyd, M. A., & Scott, D. A. (2003). Concurrent validity of written and OSCE components of the Canadian dental certification examinations. *Journal of Dental Education, 67*, 896–901.

*Graham, R., Bitzer, L. A. Z., & Anderson, O. R. (2013). Reliability and predictive validity of a comprehensive preclinical OSCE in dental education. *Journal of Dental Education, 77*(2), 161–167.

*Grand'Maison, P., Brailovsky, C. A., & Lescop, J. (1996). Content validity of the Quebec licensing examination (OSCE). Assessed by practicing physicians. *Canadian Family Physician, 42*, 254–259.

*Haque, R., Abouammoh, M. A., & Sharma, S. (2012). Validation of the Queen's University Ophthalmoscopy Objective Structured Clinical Examination Checklist to predict direct ophthalmoscopy proficiency. *Canadian Journal of Ophthalmology, 47*(6), 484–488.

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal, 1*, 447–451.

Hodges, B. (2003). Validity and the OSCE. *Medical Teacher, 25*, 250–254.

*Hodges, B., Regehr, G., Hanson, M., & McNaughton, N. (1998). Validation of an objective structured clinical examination in psychiatry. *Academic Medicine, 73*, 910–912.

Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: An introduction to generalizability theory. *The Counseling Psychologist, 27*, 325–352.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219–230.

*Hull, A. L., Hodder, S., Berger, B., Ginsberg, D., Lindheim, N., Quan, J., & Kleinhenz, M. E. (1995). Validity of three clinical performance assessments of internal medicine clerks. *Academic Medicine, 70*, 517–522.

Kane, M. T. (1982). The validity of licensure examinations. *American Psychologist, 37*, 911–918.

Lawson, D. M. (2006). Applying generalizability theory to high-stakes objective structured clinical examinations in a naturalistic environment. *Journal of Manipulative and Physiological Therapeutics, 29*, 463–467.

*Lee, M., & Wimmers, P. F. (2011). Clinical competence understood through the construct validity of three clerkship assessments. *Medical Education, 45*(8), 849–857.

*Martin, I. G., & Jolly, B. (2002). Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first year clinical year. *Medical Education, 36*, 418–425.

*Matsell, D. G., Wolfish, N. M., & Hsu, E. (1991). Reliability and validity of the objective structured clinical examination in paediatrics. *Medical Education, 25*, 293–299.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: American Council on Education/MacMillan.

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

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research, 45*, 35–44.

*Park, R. S., Chibnall, J. T., Blaskiewicz, R. J., Furman, G. E., Powell, J. K., & Mohr, C. J. (2004). Construct validity of an objective structured clinical examination (OSCE) in psychiatry: Associations with the clinical skills examination and other indicators. *Academic Psychiatry, 28*, 122–128.

*Petrusa, E. R., Blackwell, T. A., & Ainsworth, M. A. (1990). Reliability and validity of an objective structured clinical examination for assessing the clinical performance of residents. *Archives of Internal Medicine, 150*, 573–577.

*Raj, N., Badcock, L. J., Brown, G. A., Deighton, C. M., & O'Reilly, S. C. (2007). Design and validation of 2 objective structured clinical examination stations to assess core undergraduate examination skills of the hand and knee. *Journal of Rheumatology, 34*, 421–424.

*Shehmar, M., Cruikshank, M., Finn, C., Redman, C., Fraser, I., & Peile, E. (2009). A validity study of the national UK colposcopy objective structured clinical examination – Is it a test fit for purpose? *BJOG, 13*, 1796–1799.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83–117.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte: Information Age Publishing.

*Takahashi, S. G., Rothman, A., Nayer, M., Urowitz, M. B., & Crescenzi, A. M. (2012). Validation of a large-scale clinical examination for international medical graduates. *Canadian Family Physician, 58*(7), e408–e417.

*Thomson, D. M. (1987). The objective structured clinical examination for general practice: Design, validity and reliability. *The Journal of the Royal College of General Practitioners, 37*(297), 149.

*Tudiver, F., Rose, D., Banks, B., & Pfortmiller, D. (2009). Reliability and validity testing of an evidence-based medicine OSCE station. *Family Medicine, 41*, 89–91.

*Vallevand, A., & Violato, C. (2012). A predictive and construct validity study of a high-stakes objective clinical examination for assessing the clinical competence of International Medical Graduates. *Teaching and Learning in Medicine, 24*(2), 168–176.

*Varkey, P., Natt, N., Lesnick, T., Downing, S., & Yudkowsky, R. (2008). Validity evidence for an OSCE to assess competency in system-based practice and practice-based learning and improvement: A preliminary investigation. *Academic Medicine, 83*, 775–780.

*Walters, K., Osborn, D., & Raven, P. (2005). The development, validity and reliability of multimodality objective structured clinical examination in psychiatry. *Medical Education, 39*, 292–298.

Weller, J. M., Bloch, M., Young, S., Maze, M., Oyesola, S., Wyner, J., et al. (2003). Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. *British Journal of Anaesthesia, 90*, 43–47.

Chapter 17
(Mis)Alignment of Medical Education Validation Research with Contemporary Validity Theory: The Mini-CEX as an Example

Debra (Dallie) Sandilands and Bruno D. Zumbo

Like all educational assessments, assessments of medical students, residents and practicing physicians must be supported by research evidence of their validity for the purposes for which they are used. Evidence for validity is the foundation upon which meaningful and defensible interpretations of assessment results are based. The strongest evidence to support defensible use of an assessment is derived from the alignment of its validation research with contemporary validity theory as described in the *Standards for Educational and Psychological Testing* (the "*Standards*", AERA et al. 1999). The *Standards* provide criteria for the evaluation of all educational and psychological tests, testing practices and the effects of test use, as well as guidelines for test developers and users about sound and ethical use of tests. Sireci and Parker (2006) reviewed court cases involving disputes about educational tests and found that typically it is issues of test validity that are challenged in court, and that testing practices that are closely aligned with the *Standards* are more likely to withstand legal challenge. Thus in high stakes testing environments such as assessment in medical education it seems particularly important to ensure that validation efforts are aligned with contemporary validity theory as expressed in the *Standards*.

Research in other areas such as psychology and general education has found that studies are not providing validity information aligned with contemporary validity theory and that some sources of validity evidence are not being investigated or reported (Cizek et al. 2008, 2010; Hogan and Agnello 2004). Therefore the purpose

D. Sandilands, Ph.D. (✉)
Faculty of Education, Department of Educational and Counseling Psychology, and Special Education, The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
e-mail: sandilan@mail.ubc.ca

B.D. Zumbo, Ph.D. (✉)
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_17,

of this research was to investigate the extent to which studies in medical education are aligned with contemporary validity theory, using the Mini-Clinical Evaluation Exercise (Mini-CEX) (Norcini and Blank 1995) as an example. We investigated studies about the Mini-CEX because it is one of the most extensively used and studied assessment tools in medical education (Kogan et al. 2009). It has been used for more than two decades to evaluate the workplace performance of medical students, residents and physicians. Mini-CEX assessment results may have significant implications for individuals, for the educational programs that train them, and for society that relies on them to provide adequate medical care. Although there is a great deal of research that investigates the Mini-CEX, to date there has been no thorough review of the extent to which the body of Mini-CEX validation research meets the recommendations and criteria set out in the *Standards* or the extent to which the *Standards'* recommended sources of validity evidence are being reported regarding the Mini-CEX.

We conducted a systematic review of Mini-CEX studies to reveal potential gaps or limitations which may guide future Mini-CEX validation research. Specifically, our research questions were:

1. To what extent are validation studies of the Mini-CEX consistent with key aspects of contemporary validity theory as outlined in the *Standards*?; and
2. To what extent have the recommended sources of validity evidence outlined in the *Standards* been reported regarding the Mini-CEX?

It is important to note at the outset that the purpose of this study was not to evaluate the Mini-CEX or the overall quality of the research about the Mini-CEX, nor was our goal to ascertain the degree to which Mini-CEX research supports its use. Rather we were interested in gaining an understanding about how well the research is aligned with current validity theory.

In the following introductory sections we provide an overview of the Mini-CEX and of contemporary validity theory as outlined in the *Standards*.

The Mini-CEX

The Mini-CEX is a direct observation assessment tool originally developed by the American Board of Internal Medicine (ABIM) to assess the clinical skills of internal medicine residents in medical encounters with patients in a broad range of situations and locations (i.e. inpatient, outpatient, or emergency room settings). It was specifically designed to cover the skills most often required by residents in real patient encounters such as medical interviewing, physical examinations, decision-making, counseling, and clinical judgment or reasoning. The Mini-CEX is administered in two parts. First, a faculty member observes a resident while the resident conducts a focused history and physical examination on a patient, and provides a diagnosis and treatment plan. Next, immediately after the patient encounter, the faculty member gives the resident formative feedback both verbally and in writing

on a Mini-CEX rating form. The Mini-CEX rating form is said to be aligned with six (US) Accreditation Council for Graduate Medical Education (ACGME) general competencies, each of which is rated on a scale from 1 to 9. There is one additional rating for "overall clinical competence". Ratings of 1 through 3 reflect unsatisfactory performance, 4 through 6 are satisfactory (but 4 is defined as "marginal"), and 7 through 9 are superior. Each Mini-CEX takes 10–20 min to complete (ABIM 2009).

Since its inception in 1995, the Mini-CEX has been adopted for a variety of assessment purposes and is now not only used in the US but also in other countries such as Canada, the United Kingdom, Australia and Argentina. It has been suggested that the Mini-CEX may be the "only evaluation method used by many residency programs to directly observe clinical skills" (Holmboe et al. 2003, p. 826). The Mini-CEX is also used to assess residents in other specialties and its use has extended to other examinee groups such as undergraduate medical students (Dewi and Achmad 2010; Hill and Kendall 2007; Hill et al. 2009; Kogan et al. 2003; Lie et al. 2010; Ney et al. 2009), practicing doctors (Sidhu et al. 2009), and international medical graduates (Nair et al. 2008). In addition to being recommended by ABIM and ACGME, its use is also recommended by other regulators and governing bodies. As examples, the Postgraduate Medical Education and Training Board in the United Kingdom recommends the use of the Mini-CEX for assessment in the postgraduate setting (Hill et al. 2009), the Mini-CEX is mandatory during dermatology specialist training in the UK (Cohen et al. 2009), and the Australian Medical Council has introduced the Mini-CEX as a workplace assessment tool for some international medical graduates (Nair et al. 2008). In addition to providing formative feedback to guide further education and training, the Mini-CEX has been used for summative purposes to make educational decisions about medical students (Hill et al. 2009) and residents (Weller et al. 2009).

Systematic Reviews of the Mini-CEX

Two studies have used systematic reviews to investigate validity evidence for direct observation assessment methods including the Mini-CEX. Kogan et al. (2009) identified 55 tools used for direct observation and assessment and investigated evidence of their validity and outcomes. They concluded that the Mini-CEX is one of few tools that has been thoroughly evaluated and that it has the strongest validity evidence of the 55 assessment tools they investigated. However Pelgrim et al. (2010) also studied multiple direct observation tools and concluded that although the validity of the Mini-CEX is supported by correlations with other assessment instruments, additional types of validity evidence are lacking.

A third systematic review conducted by Hawkins et al. (2010) focussed specifically on the Mini-CEX and analyzed validity evidence within the framework of a validity argument (Kane 1992). Hawkins et al. (2010) found that there are relatively few studies of the Mini-CEX, the studies that do exist have variable designs that

present conflicting results, and it is "difficult to separate problems with the method from gaps and limitations in the research conducted to date." (p. 1495)

These three systematic reviews present conflicting views of the state of Mini-CEX validity research and evidence. Taken together, they raise questions about the degree and types of validity evidence that may support use of Mini-CEX scores and they highlight the need to examine potential gaps and limitations in the Mini-CEX validation research. As we noted, one way of doing this is to examine the degree to which the body of Mini-CEX validation research is aligned with contemporary validity theory.

Contemporary Validity Theory

The *Standards* (AERA et al. 1999) define validity as follows:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated.
>
> Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation refers to the construct or concepts the test is intended to measure. (p. 9)

Two aspects of the current view of validity require particular emphasis for the purposes of this paper. First, evidence for validity ought to be of highest priority to test developers, users and researchers because validity is *the most fundamental consideration* in developing and evaluating tests. In the contemporary view of validity other evidence such as reliability evidence contributes to validity and is necessary but insufficient for defensible use of test scores. Therefore validity evidence is required in addition to evidence for other characteristics such as reliability, feasibility or utility (the latter two being often reported in medical education literature).

Second, since validity pertains to interpretations and uses of test scores and not to tests themselves, validation efforts should be focussed on the proposed interpretations and uses of test scores and should begin by clearly specifying what the interpretations and uses are. In a contemporary view validation efforts consist of two inter-related arguments: an interpretive argument and a validity argument (Cronbach 1988; Hubley and Zumbo 1996; Kane 1992, 2001, 2013; Messick 1989). The interpretative argument specifies the proposed interpretations and intended uses of the test or assessment scores by identifying inferences and assumptions that flow from them, while the validity argument systematically evaluates the interpretive argument. When a particular assessment or test is used

in more than one setting for more than one application, the inferences and assumptions may change and the evidence required to support them may change. Nonetheless, validation will involve the specification (the interpretive argument) and evaluation (the validity argument) of the proposed interpretations and uses of the scores. Thus, claims about validity of test interpretations and uses are claims that the interpretive argument, the inferences and the assumptions are logical and plausible in the application in which the scores are being used (Kane 2006, 2013).

As set out in the *Standards*, specifying the interpretation begins with adequately defining the construct being measured. A construct is a broad term for the concept or characteristic a test is designed to measure, and the purpose of a test is to make inferences from test scores to unobservable constructs such as knowledge, ability, aptitude or competence. All tests should be construct-referenced because the interpretation of the construct is the foundation for the score-based inferences that arise from test use (Messick 1989). Test use and validation must proceed by clearly and thoroughly defining the construct being measured. Simply naming or labelling the construct is insufficient because the same name or label can be applied to different constructs – a common name does not automatically imply a common construct (Reckase 1998). As an example, the construct of "clinical competence" takes on different meanings when used by different parties or in different settings. Attempts to validate assessments of clinical competence should begin with a clear understanding of what is meant by clinical competence in the setting in which the assessment instrument will be used. Once the construct and proposed interpretation and inferences have been identified, evaluation through the use of a validity argument proceeds by developing empirical evidence, examining relevant literature, and/or conducting logical analyses.

Sources of Validity Evidence

The contemporary view of validity and validation requires validity evidence to be integrated from multiple sources to develop the validity argument that supports intended uses and interpretations of scores and to rule out threats to validity (Messick 1989, 1994). The *Standards* outline five sources of validity evidence that should be investigated for these purposes.

Evidence Based on Test Content

Evidence for validity can be found by analyzing the relationship between the test content and the construct intended to be measured. Sireci (1998) noted that content validity involves four commonly-accepted elements: domain definition (the conceptual and operational definitions of the construct); domain representation (match between a test and the domain definition); domain relevance (relevance of items to the content domain); and appropriate test construction procedures. Evidence based

on test content can be sought through logical or empirical analyses, including the use of subject matter experts to examine the theoretical relationship between the construct and the test content, write test items, and review item specifications, test blueprints and documentation.

Evidence Based on Response Processes

"Response processes" refers to the detailed characteristics of performance or response actually engaged in by examinees or examiners during the assessment event. Evidence based on response processes provides information about the fit between the construct and the cognitive processes engaged in during a test. For example, in a test of clinical reasoning, evidence would be required to determine whether examinees are actually using clinical reasoning skills (as opposed to perhaps following a memorized pattern of response). Evidence based on response processes can be gathered by questioning test-takers or examiners about their strategies or responses through the use of surveys, interviews, or think-aloud procedures and expert review (Miller and Linn 2000).

Evidence Based on Internal Structure

Internal structure refers to relationships between items or parts of a test. Information about a test's internal structure can reveal how closely the test conforms to the construct of interest. For example, if a test is intended to measure a unidimensional construct, then evidence of structural unidimensionality would support the relationship between the test and the construct, or if the construct is thought to be composed of several components, then multidimensionality in the test's internal structure would support that. Methods of gathering evidence based on internal structure include examining the factor structure of the data through confirmatory factor analysis, and conducting differential item functioning analyses to determine whether test items may behave differently for subgroups of examinees.

Evidence Based on Relations to Other Variables

Evidence based on relationships with other variables provides information about the extent to which the relationships are consistent with the intended construct. *Convergent validity evidence* is gathered by examining relationships between the test scores and other measures that are intended to assess theoretically-similar constructs, whereas *discriminant validity evidence* is drawn by examining relationships with measures intended to assess theoretically-different constructs. According to the *Standards*, group membership variables are relevant if the theory underlying the test use suggests that group differences should be present. For example, studies that show that scores are higher for more experienced examinees than for less

experienced examinees (or for instructed versus non-instructed examinees) provide convergent validity evidence because there is a theoretical basis for expecting score differences between the groups. *Test-criterion validity evidence* examines how accurately test scores predict a criterion performance where the criterion variable is an attribute or outcome of interest. A concurrent test-criterion study collects data from the predictor and criterion measures at approximately the same time, whereas in a predictive test-criterion study the criterion scores are obtained after the predictor scores. *Validity generalization* evidence refers to the degree to which evidence of validity based on test-criterion relations can be generalized to a new situation, for example through the use of meta-analysis. Evidence based on relations to other variables can be assessed through experimental and correlational studies, or through a multitrait-multimethod matrix approach (Campbell and Fiske 1959).

Evidence Based on Consequences of Testing

Although there is debate on this topic, evidence about the intended and unintended consequences of test use is currently required by the *Standards*. Therefore it is important to investigate whether intended consequences are occurring as anticipated, or whether *un*intended consequences may be occurring. For example, when a claim is made that a formative assessment has a positive impact on learning (such as the case of the Mini-CEX where a critical component of the assessment is the provision of feedback to examinees for the purpose of improving their performance), the validation process should question whether the positive impact is being realized.

There has been some deliberation in the literature as to whether all types of validity evidence are required for all types of assessments. The current position expressed in the *Standards* is that some sources of evidence will be especially important to evaluate in a given case, yet strong evidence from one source does not diminish the need for evidence from other sources. Therefore evidence from all five sources should be found within the body of Mini-CEX research, although they may be found to varying degrees.

Method

We conducted a search for English language literature published between January 1995 (the year in which the Mini-CEX was first introduced) and December 31, 2012 in Academic Search Complete, CINAHL, Education Research Complete, ERIC, MEDLINE, and PsychINFO. The search terms used were "Mini-Clinical Evaluation Exercise" or "Mini-CEX" and "valid*" (to capture valid, validity and validation) in all text. From this initial search we removed duplicates and excluded publications if they: (1) were not primary research, or were summaries, reviews,

interpretations or critiques of prior research; (2) did not investigate aspects of the Mini-CEX (for example, articles whose main purpose was to investigate other assessment tools but also mentioned the Mini-CEX); or (3) were editorials, letters to the editor, or conference abstracts. In addition, we examined the references in review articles to ensure the search was as comprehensive as possible.

To determine whether the main intent of each study was to present validity evidence (i.e., is the study a validity study of the Mini-CEX?), we coded whether any of the words "valid", "validity" or "validation" appeared in the title, abstract or key words and descriptors pertaining to the study. If they did we coded the study as a "validity study" and if not we coded the study as a "non-validity study".

To address the first research question regarding the extent to which the validity studies present views of validity that align with the contemporary view of validity theory, we coded whether each validity study: (1) presented a definition of validity similar to the *Standards*; (2) made reference to either the *Standards* or to contemporary validity theorists (such as those that would be taught in an introductory validity course); (3) identified and defined the construct being assessed; (4) presented a view of validity as a characteristic of Mini-CEX scores and inferences rather than as a characteristic of the Mini-CEX; (5) described the use of the Mini-CEX (for example, described the population being assessed in terms of their level of education and specialty where appropriate, the setting in which the assessment was taking place and whether the Mini-CEX scores were intended to provide formative or summative assessment information); and (6) described the intended interpretation and inferences to be drawn from Mini-CEX assessment results.

To address the second research question about the extent to which the recommended sources of validity evidence outlined in the *Standards* has been reported regarding the Mini-CEX, we coded the sources or types of validity evidence reported in the validity studies. To allow a comparison between the validity perspective taken in the studies and the validity perspective of the *Standards* and to investigate whether the sources of validity evidence being reported were aligned with sources of validity evidence in the *Standards* we re-coded the type of evidence reported in the studies as it would be reported according to the *Standards* framework. In addition, if validity evidence was presented in the non-validity studies we coded it also according to the *Standards* framework. This allowed us to fully address our second research question and determine the extent to which all recommended sources of validity evidence have been reported in all published studies of the Mini-CEX regardless of whether the studies were presented as validity studies of the Mini-CEX or not.

For both validity and non-validity studies we coded other measurement characteristics that were reported such as reliability, feasibility, utility and acceptability. Further, we coded the types of reliability evidence reported (including alternate forms, test-retest, internal consistency, scorer consistency, G-theory reproducibility, standard errors of measurement, or item response theory test information function).

All coding was carried out by the first author. In order to investigate accuracy of the coding procedure we calculated inter-rater reliability. Another researcher familiar with medical education research and contemporary validity theory coded 6 - randomly-selected studies. First, we explained the purpose of this study and reviewed the coding sheet with her. She then coded the studies independently and without knowledge of the first author's coding results.

Results

After excluding articles that did not meet the inclusion criteria as set out above, 43 articles were included in this study. A list of the included studies is attached as Appendix. Thirteen of the 43 included studies appeared to be positioned as Mini-CEX validity studies and they comprise the validity studies group. That is, 13 studies investigated the properties of the Mini-CEX and used the word "valid", "validity", or "validation" in the title, abstract or key words/descriptors pertaining to the study. The remaining 30 studies comprise the non-validity studies group.

Figure 17.1 shows the distribution of all included studies according to the year they were published. The first validity study of the Mini-CEX was published in 2002, 7 years after its inception. The majority of validity and non-validity studies have been published since 2006.

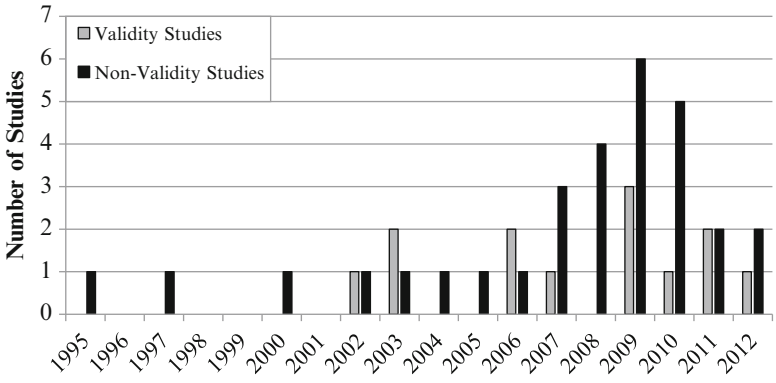

Fig. 17.1 Number of Mini-CEX studies published per year

Validity Studies' Alignment with Contemporary Validity Theory

The results of coding the 13 validity studies to determine their alignment with contemporary validity theory are summarized as follows.

None of the validity studies presented a definition of validity similar to the *Standards* although one defined "construct validity". None of the validity studies made reference to the *Standards* or to validity theorists directly, although one validity study cited an article that summarizes the *Standards* and the contemporary view of validity theory. Two validity studies provided limited (one or two sentences) definitions of the construct intended to be assessed and one study provided a reference to documentation where the construct was defined. Ten of the validity studies did not define the construct being assessed. Twelve of the 13 validity studies named a construct: 3 were reported as "competence", 4 as "clinical skills" and 5 as "clinical competence". Most validity studies named the skills that were assessed (such as history taking or physical examination) however none referred to any theoretical relationship between the skills assessed and the construct. Five validity studies clearly characterized validity as a property of the test, 5 as a property of scores or inferences, and 5 were unclear.

Figure 17.2 shows the uses of the Mini-CEX reported in the validity studies broken down by educational level, medical specialty, and assessment type. This figure reveals that for the most part the settings in which the Mini-CEX has been studied have been reported in the validity studies. As can be expected from the history of the Mini-CEX, most validity studies have investigated its use in internal medicine residencies as a form of formative assessment, although validity evidence has also been gathered for other uses and in other settings. Please note that some studies reported more than one use therefore the totals add up to more than the number of studies.

We also coded whether each validity study described how the Mini-CEX scores were to be interpreted and the inferences to be drawn from them in the particular setting of the study. No validity study specifically described the interpretation and

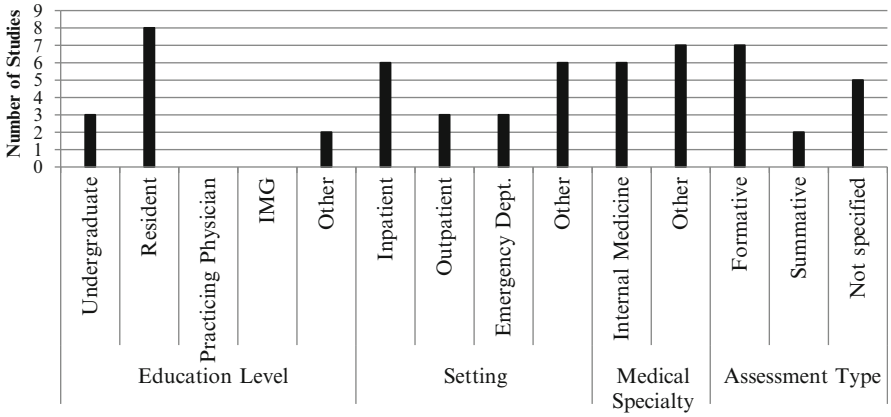

Fig. 17.2 Uses of the Mini-CEX reported in the validity studies

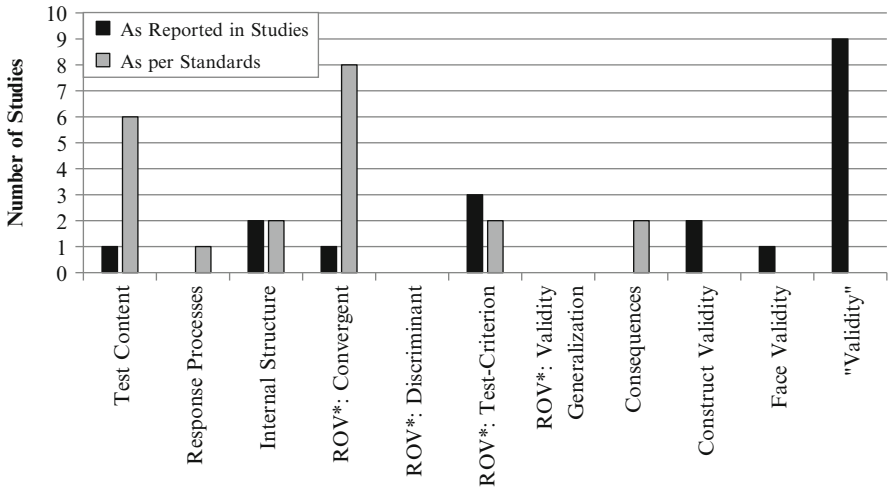

Fig. 17.3 Sources of validity evidence in the validity studies (**ROV* Relations to other Variables)

inferences that were to be drawn from the Mini-CEX scores in the particular setting of the study. Although a few studies touched on the issue, in most studies it was implicit that simply stating whether the assessment was formative or summative was sufficient to deduce whatever inferences were to be drawn.

Sources of Validity Evidence Reported in the Validity Studies

Figure 17.3 shows the sources of validity evidence as reported in the validity studies, and contrasts how the sources of validity evidence were presented in the validity studies with how the same evidence would be framed within the *Standards* framework.

Three of the 13 validity studies presented validity evidence similarly to the *Standards*; however, as can be seen in Fig. 17.3, there are considerable differences between study perspectives and *Standards'* perspectives as to sources of validity evidence in the remaining studies. Of the 9 studies that presented unspecified sources of validity evidence (i.e. evidence was referred to simply as "validity"), 4 presented evidence based on convergent relations to other variables, 4 presented test content validity evidence, 1 presented response process validity evidence, 1 presented test criterion evidence, and 2 presented evidence related to consequences of testing. In addition, 2 studies that presented construct validity evidence and 2 that presented criterion evidence were recoded as presenting evidence based on convergent relations to other variables. None of the validity studies presented evidence related to discriminant relations with other variables or validity generalization. Please note that the total number of sources of validity evidence presented is greater than the total number of validity studies because some studies presented more than one type of validity evidence.

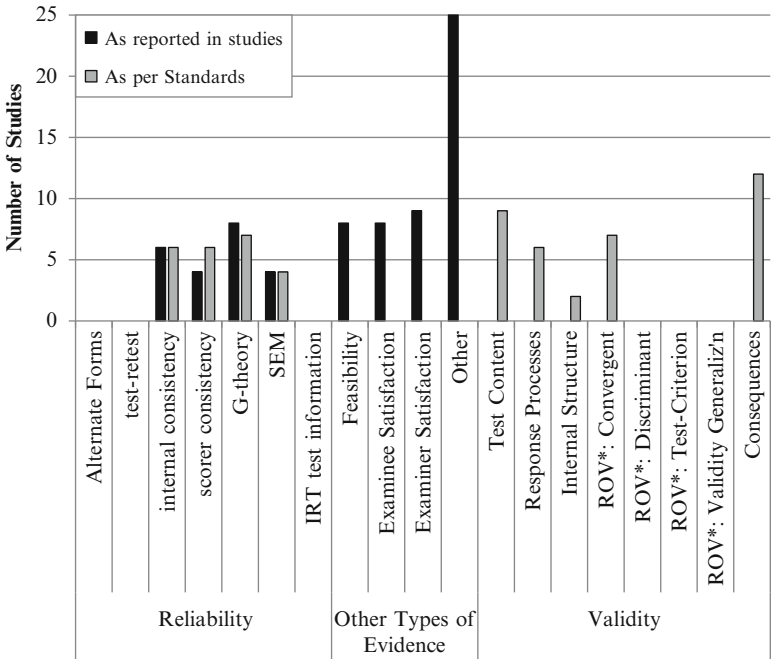

Fig. 17.4 Types of evidence presented in the non-validity studies (**ROV* Relations to other Variables)

Other Types of Evidence Reported in the Validity Studies

Of the validity studies that presented reliability evidence, five presented internal consistency evidence, three presented scorer consistency evidence, five presented generalizability theory reproducibility evidence, and two presented standard error of measurement (SEM). In addition, five of the validity studies presented feasibility evidence, four presented evidence of examinee satisfaction and four presented evidence of examiner satisfaction.

Types of Evidence Presented in the Non-validity Studies

Types of evidence presented in the 30 non-validity studies are shown in Fig. 17.4 which also contrasts how evidence was presented in the non-validity studies with how the same evidence would be framed according to the *Standards*. Most (22) of the non-validity studies reported reliability evidence and many also reported feasibility, examinee satisfaction, and examiner satisfaction. Twenty-five of the non-validity studies reported a variety of other properties of the Mini-CEX.

Examples of terms used to describe other properties were utility, accuracy, psychometric characteristics, use, acceptability, and influence on feedback. Since most studies reported more than one type of evidence the total number is greater than 30.

As in the case of the validity studies, the evidence presented for the non-validity studies would be classified differently when viewed from the perspective of the *Standards*. For the most part, reliability evidence has been framed in the studies similarly to the way it would be framed according to the *Standards* as evidenced by the similar patterns for reliability in Fig. 17.4. However, also shown in Fig. 17.4, a considerable amount of validity evidence was presented in the non-validity studies yet was not identified in the studies as validity evidence. For example, one study investigated the Mini-CEX in terms of its educational impact, the factors that influence examiner scoring decisions, and its effects on the relationship between examiner and examinee (amongst other things). According to the *Standards* these types of investigations provide information about validity such as evidence related to response processes and consequences of testing. The main types of validity evidence presented in the non-validity studies were validity evidence based on test content, response processes, convergent relations to other variables, and consequences of testing.

To What Extent Have the Recommended Sources of Validity Evidence Outlined in the Standards *Been Reported Regarding the Mini-CEX?*

Figure 17.5 shows all sources of validity evidence stemming from the 43 validity and non-validity studies combined, categorized as per the *Standards*. This figure reveals that the combined Mini-CEX research efforts (when conceptualized aligned with contemporary validity theory) have focussed predominantly on validity evidence based on test content, response processes, convergent relations to other variables, and consequences of testing. To date, the body of Mini-CEX validation research does not provide evidence based on discriminant relations to other variables or validity generalization.

Inter-rater Agreement on Coding of the Studies

Six randomly selected studies were rated by an independent rater to investigate accuracy of the coding procedure used by the first author of the study. The independent rater and first author were in agreement on 381 of the 438 total data points on the coding sheets for the 6 studies, representing 87 % inter-rater agreement. Differences were discussed and reviewed until agreement was reached.

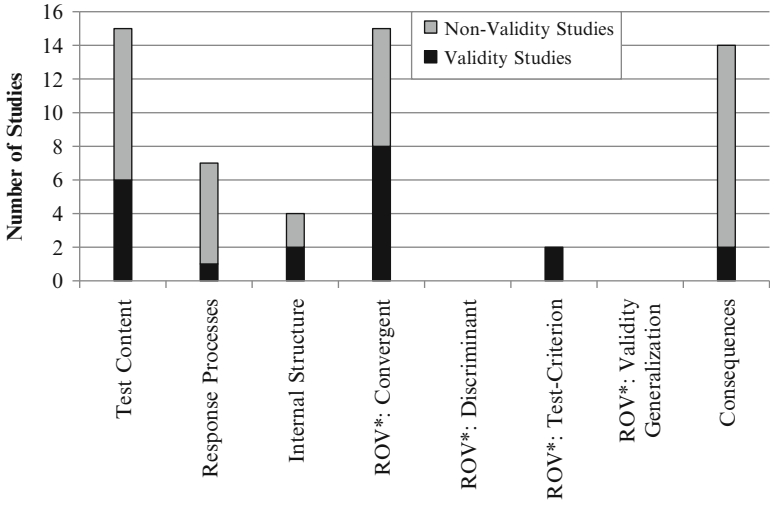

Fig. 17.5 Sources of validity evidence investigated in the validity and non-validity studies (**ROV* Relations to other Variables)

Discussion and Conclusions

This study is the first study to examine the extent to which research that investigates validity evidence for the Mini-CEX conforms to contemporary validity theory and meets the recommendations and criteria set out in the *Standards for Educational and Psychological Testing* (AERA et al. 1999). It is not the intent of this research to assess or comment on the quality of the research reviewed, rather to understand and report on the validity perspective taken in the body of literature regarding Mini-CEX and to determine which sources of validity evidence have been investigated. The results provide interesting findings about the manner in which validity is conceptualized and presented in the Mini-CEX literature and point to gaps and limitations in the research.

This study provides evidence that the body of validity research of the Mini-CEX is not fully aligned with contemporary validity theory because it does not place emphasis on the proposed interpretations and uses of test scores and on the theoretical relationship between the test and the construct being assessed. As stated in the *Standards*, validation efforts should begin with a clear definition of the proposed interpretation of scores which refers to the construct intended to be measured, together with a rationale that connects the interpretation to the proposed use of the scores. Results of the current study indicate that these first steps in the validation process for the Mini-CEX have not yet been taken: most of the validity studies investigated in this research do not provide a definition of the construct being assessed or a theoretical rationale to guide the interpretation of Mini-CEX scores.

Although most of the validity studies adequately provide contextual information about their local use of the Mini-CEX (such as the education level of the examinees, the setting in which the Mini-CEX was administered and the medical specialty that was being assessed), there was very little information about how the scores were intended to be interpreted and used. Although a few studies touched on the issue, in most studies it was implicit that simply stating whether the assessment was formative or summative was sufficient to deduce the inferences to be drawn. For example, one study reported that the Mini-CEX evaluations did not contribute to final grades of the examinees but the actual interpretations and uses of the assessments were not stated. The findings of this study confirm those of Hawkins et al. (2010) who noted that a lack of attention in the Mini-CEX validation literature to Mini-CEX use and score interpretations is a concern.

Of the studies that were presented as validity studies, none provided a definition of validity and many framed validity evidence differently than it would have been if it were aligned with the *Standards*. Only five of the validity studies characterized validity as a property of test scores or inferences. The remaining validity studies were either unclear in their position or explicitly referred to validity as a property of the Mini-CEX. For example, phrases such as "the Mini-CEX has construct validity" or "the validity of the Mini-CEX" were frequently observed in the studies. As early as the 1974 edition of the *Standards* it was considered incorrect to use the unqualified phrase "the validity of the test" (Sireci 2009) yet the results of this study point to evidence that this terminology and characterization of validity still exists in the body of research about the Mini-CEX.

This study also provides evidence about which sources of validity evidence have been reported in the Mini-CEX literature. Most studies to date have focussed on evidence based on convergent relations to other variables, test content, and consequences of testing. Few have focussed on response processes, internal structure, and test-criterion, and no studies have investigated validity evidence based on discriminant relations to other variables or validity generalization. Much of the validity evidence has arisen from studies that were not presented as having validity as their major focus and some validity evidence has been presented using other terminology such as feasibility, utility or acceptability with no connection being made to validity or validity theory. These findings support those of Pelgrim et al. (2010) who reported that few sources of validity evidence have been addressed in Mini-CEX research. They also support the findings of Hawkins et al. (2010) who found gaps and limitations in Mini-CEX validation research.

One way in which the Mini-CEX validation research is aligned with contemporary validity theory is that it is an ongoing endeavour with much research activity over the last 5 years. This practice is aligned with the *Standards* which set out that validation is a continual process and that as new uses of an assessment tool arise (as they have in the case of the Mini-CEX), research should continue to investigate sources of validity evidence associated with new use.

Implications of Findings and Suggestions for Future Research

The findings that sources of validity evidence are conceptualized differently in the published Mini-CEX validation literature than in the *Standards* and that many of the published studies presented validity evidence outside of a validity framework have implications for researchers and for journal editors. Future research could focus on enhancing researcher awareness and rectifying misunderstandings about what to report as validity evidence and how to report it. Journal editors may consider setting clear inclusion and exclusion guidelines and strengthening the peer review process for studies that investigate psychometric properties of assessments such as the Mini-CEX. Further, we found that not all studies that present validity evidence have any form of the word "valid" in their title, abstract, key words or descriptors thus making it difficult for future researchers to find the validation research that does exist. Researchers and journal editors may address this shortcoming to ensure that all future validation research will be readily accessible through typical search strategies and thus play an important role in disseminating key validity information.

The results of this study also have implications for Mini-CEX users such as medical education programs and governing bodies that set policy that recommends or mandates its use. They should be aware of the gaps in the research and degree of alignment or lack of alignment with the *Standards* and carefully consider the extent to which the existing literature supports their recommendations or the inferences to be drawn from their particular use of Mini-CEX thus ensuring that their recommendations and uses are defensible. As noted in the introduction, validation research that is closely aligned with the *Standards* most strongly supports defensible use and interpretations of test scores (Sireci and Parker 2006).

Perhaps the most important implications from this research derive from the finding that to date Mini-CEX validation research neither provides a *theoretical* rationale for score interpretation and use based on a clearly-understood construct nor clearly elucidates the inferences to be drawn from Mini-CEX use. This finding leads us to conclude that the body of Mini-CEX validation research as a whole currently represents a "weak program" of validation research (Cronbach 1988), that is, one that presents validity evidence without reference to theoretical underpinnings and often relies on data that is easily or readily available as opposed to data that is relevant (Kane 2001). Further, as noted by Kane, as early as the 1970s there was concern about the ease with which opportunistic validity evidence could be presented without stating a proposed interpretation or evaluating the reasonableness of the interpretation. In other words, the two key elements of the validation process (a clearly-stated interpretive argument and a validity argument which evaluates it) are deficient in a weak program of validation. A strong theory-driven program of research which will assure scientific and disciplined enquiry (Zumbo 2009) requires multiple strands of evidence some based on statistical analyses and some based on theory (Sireci 2009).

When a field or area of research is inhibited by the absence of well-defined theory about the construct a strong program of validity research will be difficult to achieve. It is important to note that work is being done which will help to develop theory about the construct being assessed by the Mini-CEX. For example, our study revealed research that provided validity evidence based on response processes which is theory-building research (see, as examples, Kogan et al. 2011, 2012; Weller et al. 2009). However, such research is being conducted outside of a contemporary validity theory framework. Indeed, our study revealed that a great deal of validity evidence related to the Mini-CEX has been presented in studies that fail to make any connection whatsoever to validity. A lack of connection from the research to validity or validity theory weakens or undermines the ability to develop a sound validity argument.

Kane (2001) draws distinctions between performance assessments of observable attributes and those of theoretical constructs and notes that clearly defined observable attributes might be validated with relatively simple interpretive arguments and clear validation strategies without reference to underlying theories about what is being assessed. However, the extent to which the intended interpretations generalize or go beyond the observations being made determines the strength of validity argument required: in the case of the Mini-CEX, if the intended interpretation extends from observed scores to more general conclusions about competence, then a strong program of validity research should be required. If not, a weaker program based on readily-available data may suffice. Regardless of whether the Mini-CEX is construed as assessing a theoretical construct or an observable attribute, future validation research may be directed at defining what is being assessed, building the *theoretical* rationale for score interpretation and clarifying the inferences to be drawn from Mini-CEX use thereby contributing to a stronger body of Mini-CEX validation research than currently exists.

Appendix: List of Included Studies

Alves de Lima, A., Henquin, R., Thierer, J., Paulin, J., Lamari, S., Belcastro, F., & Van der Vleuten, C. P. M. (2005). A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training. *Medical Teacher*, *27* (1), 46–52.

Alves de Lima, A., Barrero, C., Baratta, S., Castillo Costa, Y., Bortman, G., Carabajales, J., Conde, D., Galli, A., Degrange, G., & van der Vleuten, C. (2007). Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. *Medical Teacher*, *29*(8), 785–790.

Alves de Lima, A. E., Conde, D., Aldunate, L., & van der Vleuten, C. P. (2010). Teachers' experiences of the role and function of the mini clinical evaluation exercise in post-graduate training. *International Journal of Medical Education*, *1*, 68–73.

Brazil, V., Ratcliffe, L., Zhang, J., & Davin, L. (2012). Mini-CEX as a workplace-based assessment tool for interns in an emergency department – Does cost outweigh value? *Medical Teacher*, *34*(12), 1017–1023. doi:10.3109/0142159X.2012.719653.

Chen, W., Lai, M.-M., Li, T.-C., Chen, P. J., Chan, C.-Y., & Lin, C.-C. (2011). Professional development is enhanced by serving as a mini-CEX preceptor. *Journal of Continuing Education in the Health Professions*, *31*(4), 225–230.

Cohen, S. N., Farrant, P. B. J., & Taibjee, S. M. (2009). Assessing the assessments: UK dermatology trainees' views of the workplace assessment tools. *British Journal of Dermatology*, *161*(1), 34–39.

Cook, D. A., & Beckman, T. J. (2009). Does scale length matter? A comparison of nine-versus five-point rating scales for the mini-CEX. *Advances in Health Sciences Education*, *14*(5), 655–664.

Cook, D. A., Dupras, D. M., Beckman, T. J., & Thomas, K. G. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *Journal of General Internal Medicine*, *24*(1), 74–79.

Cook, D. A., Beckman, T. J., Mandrekar, J. N., & Pankratz, V. S. (2010). Internal structure of mini-CEX scores for internal medicine residents: Factor analysis and generalizability. *Advances in Health Sciences Education*, *15*(5), 633–645.

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Medical Education*, *45*(6), 560–569.

Dewi, S. P., & Achmad, T. H. (2010). Optimising feedback using the mini-CEX during the final semester programme. *Medical Education*, *44*(5), 509–509.

Durning, S. J., Cation, L. J., Markert, R. J., & Pangaro, L. N. (2002). Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Academic Medicine*, *77*(9), 900.

Fernando, N., Cleland, J., McKenzie, H., & Cassar, K. (2008). Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments. *Medical Education*, *42*(1), 89–95.

Hatala, R., Ainslie, M., Kassen, B. O., Mackie, I., & Roberts, J. M. (2006). Assessing the mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Medical Education*, *40*(10), 950–956.

Hauer, K. E. (2000). Enhancing feedback to students using the Mini-CEX (Clinical Evaluation Exercise). *Academic Medicine*, *75*(5), 524.

Hill, F., & Kendall, K. (2007). Adopting and adapting the mini-CEX as an undergraduate assessment and learning tool. *The Clinical Teacher*, *4*(4), 244–248.

Hill, F., Kendall, K., Galbraith, K., & Crossley, J. (2009). Implementing the undergraduate mini-CEX: A tailored approach at Southampton University. *Medical Education*, *43*(4), 326–334.

Holmboe, E. S., Huot, S., Chung, J., Norcini, J., & Hawkins, R. E. (2003). Construct validity of the miniClinical Evaluation Exercise (miniCEX). *Academic Medicine*, *78*(8), 826.

Holmboe, E. S., Yepes, M., Williams, F., & Huot, S. J. (2004). Feedback and the Mini Clinical Evaluation Exercise. *Journal of General Internal Medicine*, *19*(5p2), 558–561.

Jackson, D., & Wall, D. (2010). An evaluation of the use of the mini-CEX in the foundation programme. *British Journal of Hospital Medicine*, *71*(10), 584–588.

Kogan, J. R., & Hauer, K. E. (2006). Brief report: Use of the Mini-Clinical Evaluation Exercise in internal medicine core clerkships. *Journal of General Internal Medicine*, *21*(5), 501–502.

Kogan, J. R., Bellini, L. M., & Shea, J. A. (2002). Implementation of the mini-CEX to evaluate medical students' clinical skills. *Academic Medicine*, *77*(11), 1156–1157.

Kogan, J. R., Bellini, L. M., & Shea, J. A. (2003). Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Academic Medicine*, *78*(10), S33–S35.

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education*, *45*(10), 1048–1060.

Kogan, J. R., Conforti, L. N., Bernabeo, E. C., Durning, S. J., Hauer, K. E., & Holmboe, E. S. (2012). Faculty staff perceptions of feedback to residents after direct observation of clinical skills. *Medical Education*, *46*(2), 201–215.

Lie, D., Encinas, J., Stephens, F., & Prislin, M. (2010). Do faculty show the "halo effect" in rating students compared with standardized patients during a clinical examination. *The Internet Journal of Family Practice*, *8*(2). Retrieved from http://www.ispub.com/journal/the_internet_journal_of_family_practice/volume_8_number_2_20/article/do-faculty-show-the-halo-effect-in-rating-students-compared-with-standardized-patients-during-a-clinical-examination.html

Malhotra, S., Hatala, R., & Courneya, C.-A. (2008). Internal medicine residents' perceptions of the Mini-Clinical Evaluation Exercise. *Medical Teacher*, *30*(4), 414–419.

Margolis, M. J., Clauser, B. E., Cuddy, M. M., Ciccone, A., Mee, J., Harik, P., & Hawkins, R. E. (2006). Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study. *Academic Medicine*, *81*(10), S56–S60.

Mitchell, C., Bhat, S., Herbert, A., & Baker, P. (2011). Workplace-based assessments of junior doctors: Do scores predict training difficulties? *Medical Education*, *45*(12), 1190–1198.

Nair, B. R., Alexander, H. G., McGrath, B. P., Parvathy, M. S., Kilsby, E. C., Wenzel, J., Frank, I. B., Pachev, G. S., & Page, G. G. (2008). The mini clinical evaluation exercise (mini-CEX) for assessing clinical performance of international medical graduates. *The Medical Journal of Australia*, *189*(3), 159–161.

Ney, E. M., Shea, J. A., & Kogan, J. R. (2009). Predictive validity of the mini-Clinical Evaluation Exercise (mCEX): Do medical students' mCEX ratings correlate with future clinical exam performance? *Academic Medicine*, *84*(10), S17–S20.

Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The Mini-CEX (Clinical Evaluation Exercise): A preliminary investigation. *Annals of Internal Medicine*, *123*(10), 795–799.

Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1997). Examiner differences in the mini-CEX. *Advances in Health Sciences Education*, *2*(1), 27–33.

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine*, *138*(6), 476–481.

Ogunbanjo, G. A. (2009). Adapting mini-CEX scoring to improve inter-rater reliability. *Medical Education*, *43*(5), 484–485.

Quantrill, S. J., & Tun, J. K. (2012). Workplace-based assessment as an educational tool. Guide supplement 31.5-Viewpoint. *Medical Teacher*, *34*(5), 417–418.

Sidhu, R. S., Hatala, R., Barron, S., Broudo, M., Pachev, G., & Page, G. (2009). Reliability and acceptance of the Mini-Clinical Evaluation Exercise as a performance assessment of practicing physicians. *Academic Medicine*, *84*(10), S113–S115.

Torre, D. M., Simpson, D. E., Elnicki, D. M., Sebastian, J. L., & Holmboe, E. S. (2007). Feasibility, reliability and user satisfaction with a PDA-based mini-CEX to evaluate the clinical skills of third-year medical students. *Teaching and Learning in Medicine*, *19*(3), 271–277.

Van Lohuizen, M. T., Kuks, J. B., van Hell, E. A., Raat, A. N., Stewart, R. E., & Cohen-Schotanus, J. (2010). The reliability of in-training assessment when performance improvement is taken into account. *Advances in Health Sciences Education*, *15*(5), 659–669.

Weller, J. M., Jolly, B., Misur, M. P., Merry, A. F., Jones, A., Crossley, J. M., Pedersen, K., & Smith, K. (2009a). Mini-clinical evaluation exercise in anaesthesia training. *British Journal of Anaesthesia*, *102*(5), 633–641.

Weller, J. M., Jones, A., Merry, A. F., Jolly, B., & Saunders, D. (2009b). Investigation of trainee and specialist reactions to the mini-Clinical Evaluation Exercise in anaesthesia: Implications for implementation. *British Journal of Anaesthesia*, *103*(4), 524–530.

Wiles, C. M., Dawson, K., Hughes, T. A. T., Llewelyn, J. G., Morris, H. R., Pickersgill, T. P., Robertson, N. P., & Smith, P. E. M. (2007). Clinical skills evaluation of trainees in a neurology department. *Clinical Medicine*, *7*(4), 365–369.

Wilkinson, J. R., Crossley, J. G., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education*, *42*(4), 364–373.

References

American Board of Internal Medicine. (2009). *Assessment tools*. Retrieved November 20, 2010, from http://www.abim.org/program-directors-administrators/assessment-tools/mini-cex.aspx

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

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

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 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*(5), 732–743.

Cohen, S. N., Farrant, P. B. J., & Taibjee, S. M. (2009). Assessing the assessments: U.K. dermatology trainees' views of the workplace assessment tools. *British Journal of Dermatology, 161*(1), 34–39.

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

Dewi, S. P., & Achmad, T. H. (2010). Optimising feedback using the mini-CEX during the final semester programme. *Medical Education, 44*(5), 509–509.

Hawkins, R. E., Margolis, M. J., Durning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-clinical evaluation exercise: A review of the research. *Academic Medicine: Journal of the Association of American Medical Colleges, 85*(9), 1453–1461.

Hill, F., & Kendall, K. (2007). Adopting and adapting the mini-CEX as an undergraduate assessment and learning tool. *The Clinical Teacher, 4*(4), 244–248.

Hill, F., Kendall, K., Galbraith, K., & Crossley, J. (2009). Implementing the undergraduate mini-CEX: A tailored approach at Southampton University. *Medical Education, 43*(4), 326–334.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*(5), 802–812.

Holmboe, E. S., Huot, S., Chung, J., Norcini, J., & Hawkins, R. E. (2003). Construct validity of the MiniClinical evaluation exercise (MiniCEX). *Academic Medicine, 78*(8), 826–830.

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.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Washington, DC: American Council on Education.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi:10.1111/jedm.12000.

Kogan, J. R., Bellini, L. M., & Shea, J. A. (2003). Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Academic Medicine: Journal of the Association of American Medical Colleges, 78*(10), S33–S35.

Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, the Journal of the American Medical Association, 302*(12), 1316–1326.

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education, 45*(10), 1048–1060.

Kogan, J. R., Conforti, L. N., Bernabeo, E. C., Durning, S. J., Hauer, K. E., & Holmboe, E. S. (2012). Faculty staff perceptions of feedback to residents after direct observation of clinical skills. *Medical Education, 46*(2), 201–215.

Lie, D., Encinas, J., Stephens, F., & Prislin, M. (2010). Do faculty show the 'halo effect' in rating students compared with standardized patients during a clinical examination? *Internet Journal of Family Practice, 8*(2), 1–1.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*(4), 367.

Nair, B. R., Alexander, H. G., McGrath, B. P., Parvathy, M. S., Kilsby, E. C., Wenzel, J., et al. (2008). The mini clinical evaluation exercise (mini-CEX) for assessing clinical performance of international medical graduates. *The Medical Journal of Australia, 189*(3), 159–161.

Ney, E. M., Shea, J. A., & Kogan, J. R. (2009). Predictive validity of the mini-clinical evaluation exercise (mcex): Do medical students' mCEX ratings correlate with future clinical exam performance? *Academic Medicine: Journal of the Association of American Medical Colleges, 84*(10), S17–S20.

Norcini, J. J., & Blank, L. L. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine, 123*(10), 795–799.

Pelgrim, E. A., Kramer, A. W., Mokkink, H. G., van den Elsen, L., Grol, R. P., & van der Vleuten, C. P. (2010). In-training assessment using direct observation of single-patient encounters: A literature review. *Advances in Health Sciences Education*. doi:10.1007/s10459-010-9235-6.

Reckase, M. D. (1998). The interaction of values and validity assessment: Does a test's level of validity depend on a researcher's values? *Social Indicators Research, 45*(1/3, Validity Theory and the Methods Used in Validation: Perspectives from Social and Behavioral Sciences), 45–54.

Sidhu, R. S., Hatala, R., Barron, S., Broudo, M., Pachev, G., & Page, G. (2009). Reliability and acceptance of the mini-clinical evaluation exercise as a performance assessment of practicing physicians. *Academic Medicine: Journal of the Association of American Medical Colleges, 84* (10), S113–S115.

Sireci, S. G. (1998). The construct of content validity. In B. D. Zumbo (Ed.), *Validity theory and the methods used in validation: Perspectives from the social and behavioral sciences* (pp. 83–117). Amsterdam: Kluwer Academic Press.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte: Information Age Publishing.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice, 25*(3), 27–34.

Weller, J. M., Jones, A., Merry, A. F., Jolly, B., & Saunders, D. (2009). Investigation of trainee and specialist reactions to the mini-clinical evaluation exercise in anaesthesia: Implications for implementation. *British Journal of Anaesthesia, 103*(4), 524–530.

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: Information Age Publishing.

Part V
Conclusions

Chapter 18
Validation Practices in the Social, Behavioral, and Health Sciences: A Synthesis of Syntheses

Juliette Lyons-Thomas, Yan Liu, and Bruno D. Zumbo

In the first half of the twentieth century, educational and psychological researchers were aware of the importance of validity, though engaged in a variety of non-uniform methods to attain and name it (Anastasi 1986). In 1954, the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* was published jointly by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education, and thus began the effort to standardize the view of validity and the guidance for validation practice in general. Since then, researchers have expanded and refined the definition of validity, and continue to do so to date. Although content, construct, and criterion-related validity had dominated as the "trinity" view of validity, Hubley and Zumbo (1996) point out that a more unitary view has gained popularity with construct validity taking the center stage. The *Standards for Educational and Psychological Testing* (AERA et al. 1999) represents the current guidance on validity and validation practices. The *Standards* list five sources of validity evidence based on: content, internal structure, relationships to other variables, response processes, and consequences.[1] Of those five types of sources,

[1] The 2014 version of the *Standards* are not yet publicly available, however, a review of the pre-publication version indicates that the 1999 emphases and structure, in the main, remains the same.

J. Lyons-Thomas
Regents Research Fund, Institute for Urban and Minority Education, Teacher's College, Columbia University, 525 West 120th Street, 112 Zankel Hall, New York 10027, NY, USA

Y. Liu
Harvard Medical School, Harvard University, 180 Longwood Avenue, 02115, Boston, MA, USA

B.D. Zumbo, Ph.D. (✉)
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_18,

evidence related to consequences, has been particularly controversial among educational assessment researchers (Moss 1998; Nichols and Williams 2009), with some even questioning its place in validity (Cizek et al. 2008, 2010). To be more precise, nearly all of the contributors to this debate agree that consequences are relevant for assessment, in a broad sense. The disagreement seems to be around whether consequences are relevant for validation, or just generally relevant to test use.

Although the *Standards for Educational and Psychological Testing* has been published in 1999 and the validity issues have been discussed in many fields and journals in the past decade, there still remain a lot of concerns and questions about whether and to what extent the current validation practice adopted in the published journal papers has followed the *Standards*. To provide a window into this issue, the previous 15 chapters collected in this volume have synthesized validity evidence and validation practice, which either present the current validation practice or reflect the change of validation practice over time. The present chapter is meant to summarize the shared findings as well as the differences found across the 15 validity synthesis chapters. An attempt is made to provide insight into where the research on validity presently stands, how it has changed from its inception, and where it is heading across a broad range of disciplines and journals in the educational, psychosocial, and health sciences domains. In terms of our meta-synthesis, emphasis is placed on the improvement and the benefits that validation-oriented research has for these domains of inquiry and the importance of engaging in it to appropriately use educational and psychological tests and measures and interpret test scores.

Data Sources and Methodology

In order to accomplish the objective set forth above, the 15 synthesis chapters from this book were compared to one another based on the information that was collected about validation practices. A common element of all of the chapters was that each examined validity evidence according to the *Standards* (AERA et al. 1999). That is, each chapter provided a numerical summary for the five sources of validity evidence based on: (a) content, (b) response processes, (c) internal structure, (d) associations with other variables, and (e) consequences. In addition, each chapter included other validity evidence relevant to their research area. For instance, some papers included a count of articles that provided face validity evidence, though other papers did not consider face validity, either because it was not regarded as part of validity evidence by the *Standards* or because it was not relevant to their purposes.

It should also be noted that despite the common theme of examining validation evidence, the system of determining which information to include varied from paper to paper. While some chapters tallied validity evidence based on how it was reported, others reported validity evidence based on the authors' own

evaluation, that is, the judgment by the authors on the validity evidence that should have been reported. For instance, a main argument from Chap. 17, (Sandilands and Zumbo), is that there is misrepresentation from many studies that purport to present validity evidence in the area of medical education. Therefore, the authors chose to report both validity evidence as it was presented in the research articles, as well as validity evidence as the authors thought it should have been reported.

Another dissimilarity among the chapters is that there was variation across domain and temporal period. Papers focused on validation practices in different areas, such as education, counseling, health, well-being, medical education, or psychology. Furthermore, some authors focused on particular instruments, while others directed their synthesis on individual journals, and two chapters even focused on specific journals within two different time periods to compare if and how validation practices have changed with the evolving concept of validity. Additionally, many but not all chapters noted whether papers cited or integrated validity theory or framework in their validation practices (e.g., AERA et al. 1999; Kane 2006; Messick 1989). Given both the unique and overlapping characteristics of the chapters, this meta-synthesis did not solely focus on numerical analysis, but also compared and contrasted the features of the synthesis chapters in a qualitative way to describe overall trends in validity research.

Major Findings from All Synthesis Chapters

The 15 synthesis chapters provide rich information about the current validation practice across a variety of disciplines and from different journals. Our review here will only focuses on the validity view adopted in the validation practice, the misconceptions frequently occurred, and most popular validity evidence as well as the most neglected validity evidence.

One of the findings was the wide acknowledgement of the importance of validity and an increase in the number of researchers trying to empirically ground the usefulness and appropriateness of the conclusions derived from the scores of the instruments. However, despite the wide-ranging acknowledgement of the importance of validity, references to the *Standards* is practically non-existent. Furthermore, many validation studies are still firmly grounded in early twentieth century conceptions that view validity as a property of the test, without acknowledging the importance of building a validity argument to support the inferences of test scores. There appears to be minimal evidence of recognition of the modern/unitary view of validity. With respect to the field of study that appears to be most in line with contemporary views of validity and validation practices, it may come as no surprise that the measurement focused journal *Educational and Psychological Measurement* was found to be the most current.

There were also some misconceptions found with respect to the types of evidence that are presented when attempting to make a validity argument. We found that although validity evidence based on relationships and comparisons with other

variables was widely reported, there seems to be some confusion across disciplines with regard to terminology and the nature of the evidence. For instance, there were misunderstandings between discriminant versus discriminative evidence and criterion-related validity evidence was sometimes presented as predictive validity evidence.

An interesting finding from the chapters has to do with evidence related to internal structure and its apparent increase over time. Both Collie and Zumbo (Chap. 7) and Shear and Zumbo (Chap. 6), which compared validity evidence in the 1950s and 1960s, respectively, to validity evidence in the 2000s, found that the number of journal papers that included internal structure evidence dramatically increased over time. While the other chapters only looked at more recent validation studies, the findings from those papers appear to support the high use of internal structure. Out of all of the categories of validity that was coded, internal structure had the highest rate. For instance, two out of the three syntheses from the Chan et al. chapter (Chap. 5) Validity Evidence and Validation Practices in Counseling: Major Journals, Mattering Instruments, and the Kuder Occupational Interest Survey (KOIS), found that evidence based on internal structure was presented in almost all (95.2 %) of the papers that were coded.

Finally, an important finding from compiling the results of the chapters is that two sources of validity evidence appear to be used rarely, if at all, across all fields. Table 18.1 displays the percentage of articles that presented evidence based on response processes and consequences for each synthesis chapter. It showed that validity evidence based on response processes and evidence based on consequences has been virtually ignored in the validation of scales, most studies showing zero percentage of reporting these two sources of evidence. From a temporal perspective, these two sources of validity evidence have remained overlooked in practice, despite the evolution of validity theory and the intense discussion of these types of evidence. The Sandilands and Zumbo synthesis (Chap. 17), which found the highest amount of evidence related to consequences, also reported that most of the studies did *not* position themselves as validity papers. As described earlier, evidence based on consequences is controversial. However, the complete lack of acknowledgement across disciplines suggests that current conceptions of validity have not yet permeated practice.

Discussion

The 15 syntheses chapters demonstrate that a number of patterns are present in current validation research across a variety of areas. Despite the changing face of validity, validation research appears to remain stagnant in the early theoretical validity framework, with the exception of the increase in evidence based on internal structure. One possible cause for this is that, between the midcentury and present day, methods of collecting evidence based on internal structure have become increasingly accessible and even required by many journals. Technology and

Table 18.1 Percentage of articles that include evidence based on response processes and consequences

Chapter	Focus of review, journal/measure	Response processes	Consequences
Chapter 3, "Reporting of Measurement Validity in Articles Published in *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*"	Journal	9.5 %	0 %
Chapter 4, "A Research Synthesis of Validation Practices Used to Evaluate the Satisfaction with Life Scale (SWLS)"	Measure	4.3 %	0 %
Chapter 5, "Validation Practices in Counseling: Major Journals, Mattering Instruments, and the Kuder Occupational Interest Survey (KOIS)" (Study 1)	Journal	0 %	0 %
Chapter 5, "Validation Practices in Counseling: Major Journals, Mattering Instruments, and the Kuder Occupational Interest Survey (KOIS)" (Study 2)	Measure	0 %	0 %
Chapter 5, "Validation Practices in Counseling: Major Journals, Mattering Instruments, and the Kuder Occupational Interest Survey (KOIS)" (Study 3)	Measure	0 %	0 %
Chapter 6, "What Counts as Evidence: A Review of Validity Studies in *Educational and Psychological Measurement*"	Journal	5 %	0 %
Chapter 7, "Validity Evidence in the *Journal of Educational Psychology*: Documenting Current Practice and a Comparison with Earlier Practice"	Journal	0 %	0 %
Chapter 8, "A Review of Validity Evidence Presented in the *Journal of Sport and Exercise Psychology* (2002–2012): Misconceptions and Recommendations for Validation Research"	Journal	2 %	0 %
Chapter 9, "The Edinburgh Postnatal Depression Scale (EPDS): A Review of the Reported Validity Evidence"	Measure	1.8 %	3.5 %
Chapter 10, "Validity Theory and Validity Evidence for Scores Derived from the Behavioural Regulation in Exercise Questionnaire"	Measure	0 %	0 %
Chapter 11, "Synthesis of Validation Practices in Two Assessment Journals: *Psychological Assessment* and the *European Journal of Psychological Assessment*"	Journal	1.8 %	0 %
Chapter 12, "Reporting of Measurement Validity in Articles Published in *Quality of Life Research*"	Journal	0 %	0 %%

(continued)

Table 18.1 (continued)

Chapter	Focus of review, journal/measure	Response processes	Consequences
Chapter 13, "Validity Evidence for a Perceived Social Support Measure in a Population Health Context"	Measure	0 %	0 %
Chapter 14, "Medical Outcomes Study Short Form-36 (SF-36) and the World Health Organization Quality of Life (WHOQoL) Assessment: Reporting of Psychometric Validity Evidence"	Measure	0 %	0 %
Chapter 15, "Reporting of Validity Evidence in the Field of Health Care: A Focus on Papers Published in *Value in Health*"	Journal	4.4 %	2.9 %
Chapter 16, "Validation Practices of the Objective Structured Clinical Examination (OSCE)"	Measure	9.1 %	4.5 %
Chapter 17, "(Mis)Alignment of Medical Education Validation Research with Contemporary Validity Theory: The Mini-CEX as an Example"	Measure	7.7 %	15.4 %

user-friendly software programs have become ubiquitous with research, and the ease with which one can perform a factor analysis or item analysis has transitioned from an arduous process to "point and click".

Perhaps the most important finding from this review is that two particular sources of validity evidence are largely ignored across disciplines, despite their addition as important sources of validity evidence in the *Standards*: response processes and consequences. Despite these findings, and the inclination to assume that one or both of these two sources of evidence do not belong in validation research, it may be prudent for future investigations to examine the underlying reasons behind this lack of evidence. One possible explanation behind the lack of evidence related to response processes is that data collection of such evidence is time consuming. Using a practice such as think aloud protocols to understand cognitive processes requires one-on-one interview sessions, transcribing, coding, and then finally analyses. Meanwhile, an aversion to addressing consequences may simply reflect the current climate of measurement research. In this area, evidence related to consequences is hotly debated, and at times discouraged. For this reason, it could conceivably be avoided by some researchers. In any case, one future direction of research lies in understanding researchers' conceptual understanding of validity, how these two sources of evidence fit with validity research, and a deeper investigation of the methodology of investigating this type of evidence.

References

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

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15. doi:10.1146/annurev.ps.37.020186.000245.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 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*(5), 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*(3), 207–215.

Kane, M. (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.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice, 28*(1), 3–9.

Chapter 19
Reflections on Validation Practices in the Social, Behavioral, and Health Sciences

Bruno D. Zumbo and Eric K.H. Chan

The volume is a high watermark for the field of assessment, testing, and measurement because, to the best of our knowledge, it is the first such project wherein such a wide variety of validation studies across the social, behavioral, and health sciences were closely examined to document validation practices. Papers published in psychology, education, epidemiology, kinesiology, medical education, educational psychology, quality of life and well-being, counseling, and patient-reported outcomes were included in this volume. The 15 research syntheses of practices provide a detailed study of the genre of validation writing – its focus, style, orientation, and structure. Of course, like all studies of published writing and genre, the papers reflect not only what the authors chose to emphasize (and how they chose to do so) but it also reflects what editors and reviewers are requiring, as well as allowing, as validation evidence.

It is important to keep in mind that in the course of conducting the syntheses some of our chapters are focused on what the validation researchers said they reported whereas others re-categorized and scrutinized what the validation researchers reported in their study. Therefore, some chapter authors reported what the validation researchers claimed to be doing whereas others scrutinized what the validation researchers claimed and, where necessary, re-categorized the validity evidence in their analysis. Readers need to keep this distinction in mind.

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_19,

Reflections on Current Practices

Before turning to more broad remarks, let us begin with very narrowly focused remarks about validation practices that stood out to us as salient. Overall, the most common sources of validity evidence that are reported are 'construct validity' and 'relations to other variables'. Two points are noteworthy about this. First, there is no universal understanding, or claim, of what constitutes 'construct validity' evidence. The most common form of construct validity evidence is a statistical investigation of the internal structure of the measure via some form of dimensional analysis such as factor analysis. This operationalization of construct validity has persisted since the 1960s with the only significant change being how the factor analyses are conducted. Contemporary methods of studying the internal structure include confirmatory factor analysis and periodically item response theory. A second point is in regard to examining relations with other variables. Hubley and colleagues (Chap. 11) provide a detailed critique in their closing section of their chapter; however, the most salient observation for us is the high frequency of convergent evidence across all of the chapters but relatively low inclusion rates of discriminant evidence. As Hubley and Zumbo (2013) noted, it is useful to think of convergent and discriminant measures as being on a continuum wherein correlations between theoretically similar measures (i.e., convergent validity) should be 'relatively high' while correlations between theoretically dissimilar measures (i.e., discriminant validity) should be 'relatively low'. This permits the researcher to better interpret the obtained validity coefficients. Therefore, when conducting a study using convergent measures, it is important to include discriminant measures for comparison, and to pre-specify the expected relative magnitude of coefficients from, or the rank order of, each of the convergent and discriminant measures. Likewise, it is unclear in contemporary practice what qualifies as a 'criterion' in criterion-related evidence. In the early parts of the last century, when tests were akin to predictive devices of behaviors, the criterion was far clearer. As Hubley and her colleagues note, in contemporary practice there are many cases in which the evidence being presented should be treated as convergent evidence rather than criterion-related evidence.

On a more general note, we found that the validation practices had a feel of being, in best light, opportunistic to, in much worse light, somewhat haphazard. Our own experiences as reviewers is that validation studies sometimes read as if the primary study was about a substantive area of research and that the authors have ancillary data that could serve as a source for validity evidence.[1] One can see now why we used the term "haphazard". The term haphazard is meant to suggest that some of the validation studies are characterized by a lack of order or planning and

[1] This remark about haphazard validation practices excludes the studies of professional testing organizations, testing and assessment divisions in government agencies, and test publishers. In the last 10 years we have observed a marked increase and interest in systematic validation plans at testing agencies and institutions.

somewhat determined by chance inclusion of some criterion variable; which explains the concern described above about the practice of investigating the relation to other variables. This feel of being "haphazard" or "opportunistic" validation studies could reflect funding and granting agency priorities, biases and proclivities of tenure and promotion committees and academic awards committees, as well as graduate training. Simply put, and in our experience, researchers are discouraged from extensive validation work either via funding priorities, editorial policies at journals, and/or academic review committees.

Our second general observation from the 15 research syntheses is that, by and large, validation studies are not guided by any theoretical orientation, validity perspectives or, if you will, validity theory. This is problematic because the activity becomes very piecemeal and unfocused. As Shear and Zumbo (Chap. 6) note, some validity theorists have stated that the current unified theory of construct validity, as described by Messick (1989) and in the *Standards*, requires unattainable or unrealistic goals (Chapelle et al. 2010; Lissitz and Samuelsen 2007; Moss 2007; Shepard 1993). As a result, as Shepard (1993, p. 429) states, "the sense that the task is insurmountable allows practitioners to think that a little bit of evidence of whatever type will suffice" (Shepard 1993, p. 429).

An alternative explanation is that the *Standards* and other descriptions of current validity theories simply lack practical guidance (Kane 2004, 2006). In this light, Chan and his colleagues (Chap. 5) call for reporting guidelines with a set of recommended items that authors should report for their validation studies. Several potential guidelines and standards are reviewed by Chan (Chap. 2). Having accepted and endorsed reporting guidelines on validity would allow the standardization of information reported in validation studies and improve the quality of the peer review process. Likewise, Chan reminds us of the important distinction between standards and guidelines and that ironically the *Standards,* which are endorsed by the American Psychological Association (APA), appear to only meet APA's own description of guidelines. Chan and colleagues (Chap. 5) note that if reporting guidelines for validation studies are established, they need to be adopted by researchers, journal editors, journal reviewers, and the broader academic community. The use of accepted reporting guidelines is associated with better quality academic publications (Cobo et al. 2011). Journal editors play an important role in the peer review process, they are therefore in the best position to promote the use of guidelines for the reporting of validity evidence. Finally, more concerted efforts are needed to expand the graduate curriculum to include courses or seminars in contemporary validity theories and practices.

In a philosophic sense the field of validity and validation is at a "pre-scientific" (a la the philosopher Thomas Kuhn) stage of development because there is not yet widely agreed upon exemplars of good validation practice. This is not to suggest that good work is not going on; but rather that it has not reached the stage of being an exemplar to guide others' validation practices. Kane's wonderfully elaborated "argument-based approach" has a lot of good advice but has seemed to evolve in to a way to "think about" validation practice rather than a series of exemplars for practice. To be fair, there is no evidence that with his argument-based approach

Kane set out to provide the Kuhnian exemplars that would necessitate a paradigm shift. Ours is just an observation that there may have been a (yet to be resolved) missed opportunity. In fact, to highlight this point of a lack of exemplars further, it is well-known that after the publication of the 1999 *Standards* a group was convened to do precisely this task: find exemplars and share them widely alongside the *Standards*. This group, all leading experts, however, could not agree on what constituted an exemplar validation that could be modeled by others. It is for this reason alone that we consider validation at a Kuhnian pre-scientific stage.

Our recommendation is that validation studies (and, ideally test developers or adaptors of tests to different languages, cultures, or contexts) need to have an explicit "validation plan" and the plan needs to be guided by some conceptual or theoretic orientation; much like practicing researchers would on a day-to-day basis. We have our own leanings to having explanation of score variation as a regulative ideal and the inclusion of consequences in an expanded model of validation (Hubley and Zumbo 2011, 2013; Zumbo 2007, 2009; Zumbo and Forer 2011; Forer and Zumbo 2011) but frankly any orientation or guidance would be adequate because it will be a guiding feature of the validation study. Although Zumbo (2007, 2009) has argued that a theory of validity is important when measurement specialists and psychometricians are developing methods for validation; in the practice of validation which "theory" or "theories" of validation one chooses to use is less important than the fact that a theory is needed. There are many to choose from: Cronbach-Kane argument based approach (Kane 2013), Embretson's (1983, 2007) approach, Zumbo's approach, or others. In fact, the basic elements of the *Standards* could be used to create a validation plan. For example, for the last 20 years the first author (Zumbo) has been presenting validation methods in his graduate courses by creating a table wherein the columns are the five sources of validity evidence from the *Standards* and the rows are the various purposes and uses of the test or measure. Along with an initial discussion of the uses of the test or measure, and a consideration of construct irrelevant variance and construct under-representation, this simple grid forces validation planning to consider a structured form of validity evidence.

In summary, as Shear and Zumbo (Chap. 6) note, the absence of guiding theories of validity is more troubling than the absence of any one particular concept of validity. In the absence of a clear guiding theory of validity, it is difficult to evaluate whether a particular program of validity research has accomplished its aims. In its essence, this absence undermines the statement in the *Standards* that validity is "the most fundamental consideration in developing and evaluating tests" (AERA et al. 1999, p. 9) because it may not be clear what exactly a concern for validity entails.

What Are Consistently Under-Represented in Validation Practices: Response Processes and Consequences

We wish to close our remarks with some reflections on what is consistently under-represented in validation practices: a concern for response processes and consequences. Several of the chapter authors spoke to explanations for why these two sources of validity evidence are under-represented (see, for example, Shear and Zumbo, Chap. 6; Hubley et al., Chap. 11; Chinni and Hubley, Chap. 4) so we will not tread over that well worn path again. Instead, we wish to highlight that this under-representation is a truly considerable missed opportunity. First, it is important to expand the evidential basis of validity to include qualitative and mixed methods evidence. Second, methods like think aloud processes or cognitive interviews are useful in unpacking how test takers, or anyone taking an assessment or measure, are responding. A paper that may be useful as an exemplar would be the extensive study by Gadermann et al. (2011). Likewise, an important program of research is being conducted by José-Luis Padilla and his colleagues on cognitive interviews (for example, Castillo-Díaz and Padilla 2013). In our opinion this sort of evidence, either cognitive interviews or talk aloud processes, suits well with Zumbo's explanatory view of validity, but it also fits well with Messick's 'substantive' evidence.

In terms of consequences, we would like to highlight the Hubley-Zumbo unified framework of validity and validation (Hubley and Zumbo 2011, 2013) as depicted in Fig. 19.1.

To read and apply the framework, one would start at the far left of the figure with theories that define the variable of interest and also explicitly articulate its proposed uses (and ideally what it should not be used for). One then moves from left to right with a clear eye for when the loops double-back in the process. As Hubley and Zumbo state, their framework is consistent with Zumbo's (2009) view of validation as an integrative cognitive judgement involving a form of contextualized and pragmatic view of explanation – wherein explanation serves as a regulative ideal. Furthermore, their framework pays greater attention to the roles of values and theory at each step of validation, types of evidence included in construct validation (see the large dashed circle at the center of the framework in Fig. 19.1), and the role of intended consequences and unintended side-effects (concepts that they more fully introduce and explicate in their paper). Unlike Messick, the leading protagonist for the role of consequences in assessment and validation, the Hubley-Zumbo framework shows that from test score meaning and inference emerge both intended social and personal consequences as well as unintended social and personal side effects of legitimate test use. And importantly consequences and side effects of legitimate test use may also influence test score meaning, inferences, and decisions, which make them relevant to the validation process. Finally, in Fig. 19.1 the fact that some of the arrows loop back in the framework is particularly important such that consequences and side effects of legitimate test use can affect the articulation of the construct. Likewise, we can see that the role of values is pervasive throughout

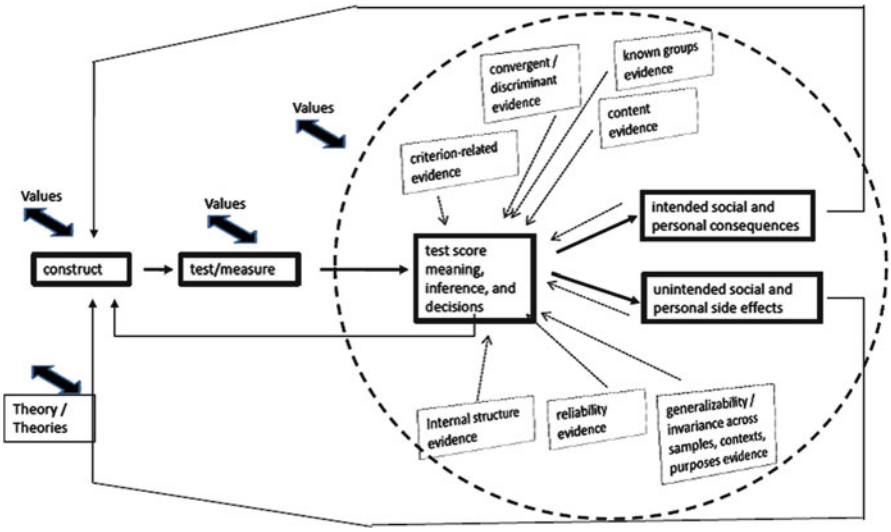

Fig. 19.1 The Hubley-Zumbo framework of validity and validation (*Note:* This figure was adapted from page 226 of Hubley and Zumbo (2011))

the framework and are related to theory/theories (broadly defined), the construct, construct validation choices, and decisions. We propose that the Hubley-Zumbo framework can guide the role and purpose of consequences in validation practice. Although it may, at first blush, be seen as a radical departure from current validation theory and practices it embodies, for the most part, contemporary thinking in the field.

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.

Castillo-Díaz, M., & Padilla, J.-L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research, 114*, 963–975.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Cobo, E., Cortés, J., Ribera, J. M., Cardellach, F., Selva-O'Callaghan, A., Kostov, B., et al. (2011). Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: Masked randomized trial. *British Medical Journal, 343*, d6783.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher, 36*(8), 449–455.

Forer, B., & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research, 103*(2), 231–265.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the satisfaction with life scale adapted for children: A focus on cognitive processes. *Social Indicators Research, 100*, 37–60.

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. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspective, 2*(3), 135–170.

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.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437–448.

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

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher, 36*(8), 470–476.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*(1), 405–450.

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/Boston: 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: Information Age Publishing.

Zumbo, B. D., & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J. A. Bovaird & K. F. Geisinger (Eds.), *High stakes testing in education: Science and practice in K-12 settings* (pp. 177–190). Washington, DC: American Psychological Association.