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.