# Validity: Challenges in Conception, Methods, and Interpretation in Survey Research

Menold, Natalja; Bluemke, Matthias; Hubley, Anita M.

Postprint / Postprint
Zeitschriftenartikel / journal article

# Editorial

## Validity

Challenges in Conception, Methods, and Interpretation in Survey Research

Natalja Menold,[1] Matthias Bluemke,[1] and Anita M. Hubley[2]

[1] GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
[2] Department of Educational & Counselling Psychology & Special Education, University of British Columbia, Vancouver, BC, Canada

Validity is *central* to measurement in the social sciences, regardless of whether diagnostic tests, survey questions, or the adequacy of interpretations of measurement outcomes are concerned. Even still, researchers developing and using measurement instruments are faced with a diversity of validity conceptions. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) include a description of validity and set of standards for validation practice. Although the *Standards* were developed in the United States, they have been shown to have worldwide impact (Zumbo, 2014). The *Standards* defined validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11) and promotes five sources of validity. While not fully endorsed by all measurement experts, as seen in several special issues devoted to this topic (e.g., Newton, 2012; Newton & Baird, 2016), this "consensus" definition provides an important reference point for developers and users of measurement instruments and a touchstone for those whose views deviate from it, be they survey researchers, test developers, or psychometricians. Markus and Borsboom's (2013) book on validity is a recent example of the interest in challenging and refining the conceptions of validity.

However, there is clearly a gap between the recommended guidelines and validation practice. Validation syntheses have shown that internal structure and relations to other variables have been the dominant sources of evidence in the literature (e.g., Zumbo & Chan, 2014). Their long-standing history in validity practice has likely been fostered by statistical and software advances that make it relatively easy to obtain this specific kind of evidence. By contrast, although evidence supporting the content of a test or measurement instrument is considered important, particularly during the development phase, it tends to be underreported. Worse still, response processes and test consequences have been sadly neglected in validation practice in the area of psychology (Hubley, 2018; Zumbo & Chan, 2014), although survey research has particularly addressed the psychology of response processes and its role in measurement errors (e.g., Tourangeau, Rips, & Rasinski, 2000). The recent publication of books on advances in response processes as a source of validity evidence in the social, behavioral, and health sciences (Zumbo & Hubley, 2017) and in second generation educational assessment (Ercikan & Pellegrino, 2017) suggest a growing interest in this source of evidence.

Maul (2017) raised concerns about traditional validation approaches with survey based self-report measures. There is still a strong need for good exemplars of validation studies and guidance for those interested in developing self-report measures and validating the inferences made from those measures. Although Kane (2006, 2013) has provided a very useful general framework with his two-step argument-based approach to validation, there is still a need for even more specific guidance. Rammstedt et al. (2015) developed validation standards for social science survey research. Organizations, such as GESIS – Leibniz Institute for the Social Sciences, have committed to promoting discussions among experts on the challenges of validation as a relevant part of scale construction and the measurement process, through regularly conducted meetings of experts on scale construction. This special issue was initiated as an outreach of the GESIS expert meeting "Advances in Scale Development in the Social Sciences: Ensuring Validity", which took place on December 1–2, 2016 in Mannheim, Germany.

The goal of this special issue is to provoke a dialogue on issues in conceptualizing validity, enhance knowledge and application of validation methods, and shed more light on challenges faced in interpretation and use of measures in psychological and social science research. These topics are particularly important given the complexity and heterogeneity not only of concepts and populations under

investigation in the social sciences, but also due to the increasing complexity of data collection processes in large scale surveys. The latter refers to, for example, mixed-mode (face-to-face, postal, internet) or mixed device surveys (on PCs or smartphones), adaptive design surveys, interviewer effects, or other situational effects.

In response to our initial call for extended abstracts for articles for this special issue on Validity in Survey Research, we were delighted, but taken aback, to receive 45 submissions! We think this high submission rate speaks to a strong interest in, and response to challenges related to, validity and validation in survey research. We received many very strong and interesting proposals but we were restricted in the number we could pursue and, after much debate, invited authors to submit full articles on the basis of abstracts that made a strong link to validity, had broad appeal or relevance to the journal's readership, and were able to provide a strong level of detail with respect to research questions, research design and results. After undergoing peer review, the special issue consists of five feature articles.

Sandra Camargo, Aura Nidia Herrera, and Anne Traynor (2018) tackle the ongoing debate surrounding conceptions of validity among leading validity theorists. Using a Delphi study, they identify issues in validity theory for which there is a consensus among the surveyed experts, and the issues on which the experts continue to disagree. Given the numerous papers and special issues that have appeared over the years on validity theory and practice, this study provides some welcome insight on points of agreement and contention. We are convinced that this article will help to initiate further studies, which evaluate conception, understanding, assessment procedures and use of results with respect to validity in social science research and thus foster further awareness and development of theoretical conceptions and validation procedures.

One challenge is to develop a theory of formative measurement and to provide a methodological basis for validation of formative scales. In contrast to the many measures that are built on a reflective measurement model, formative scales cannot rely on evidence of internally consistent item sets. The article by Keith A. Markus (2018) brings our thinking a step forward in conceptualizing issues related to formative scales. The author discusses a range of essential questions related to scale construction, item analysis and assessment of validity of formative scales. Research questions include: "What are the benefits of formative scales and under what conditions would one deliberately construct a formative scale?" or "How might one conceptualize bias in the context of formative scales?" (Markus, 2018, p. 158). In sum, Markus challenges readers to abandon restrictive notions of measurement, the flawed notion of conceptual unity, and the conflation of item scores with attributes if they want to develop a scale theory for formative scales.

Another challenge in state-of-the-art measurement is to take into account the heterogeneity of populations under investigation. If several populations are diverse with respect to their educational and social context, or if subgroups within the same population show different biased reactions to item content, this can hamper score comparability and invoke measurement bias to an unknown degree. This bias can be observed and tested in form of differential item functioning (DIF), meaning that respondents with the same true value will provide different observed scores. In this special issue, Anne M. Gadermann, Michelle Chen, Scott Emerson, and Bruno D. Zumbo (2018) compare three different methods for evaluating DIF in ordinal response scales and discuss their use in social science research. One implication is that researchers are invited to pay more attention to, and implement, DIF analyses regularly. Especially for cross-cultural surveys with rather few (say, less than 5) response options, there is a chance to accurately and critically examine the statistical comparability of scores across countries or other groups of population.

Whereas the aforementioned DIF analyses are applied to existing measurement instruments *after* data collection, comparability across different groups might already be addressed during the scale development phase and piloting of measurement instruments. To ensure comparability at an earlier research stage, Martin Schultze and Michael Eid (2018) introduce an automatic procedure for the development of short, economic scales, taking measurement invariance into account. The authors present an adaptation of the MAX-MIN Ant-System algorithm and demonstrate the possibilities for selecting suitable items to arrive at short, but reliable, cross-culturally comparable multi-item scales. This approach will be useful when adapting psychological scales for cross-cultural measurement in social science population surveys, which is becoming more and more important (Rammstedt & Beierlein, 2014).

Finally, Esther Beierl, Moritz Heene, and Markus Bühner (2018) evaluate goodness-of-fit indices for Confirmatory Factor Analysis (CFA) proposed by Hu and Bentler (1999) and widely adopted in psychological, behavioral and social science research. In a simulation study, the authors investigate the sensitivity of common fit indices with respect to model misspecifications when testing the dimensionality of a measure. An assessment of dimensionality of a measure is a central task when establishing factorial validity. The results provided by the authors point out that, when using cut-off heuristics for goodness-of-fit, the complexity of the measure should be taken into account. Rather than limit evaluation of the results by the use of goodness-of-fit indices, the authors encourage paying attention to other

results of model estimation, particularly the factor loadings. This study also demonstrates that more research is needed to define adequate goodness-of-fit indices for multidimensional measures or complex factor structures.

Overall, the articles in this special issue make innovative and important contributions to address questions of validity conceptualization, validation process, and use of measurement results. All contributions are directly relevant to the practice of scale construction in different areas of behavioral and social science research. With the articles included in this special issue, the guest editors hope to foster both substantial theoretical evolution and development of innovative analytical approaches of validity evaluation. It is our hope to initiate and encourage deep discussions by and among methodologists, validity theorists, and researchers with respect to the key questions in validity assessment as well as the interpretation and further use of measurement outcomes.

## References


American Educational Research Association; American Psychological Association; National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Beierl, E., Heene, M., & Bühner, M. (2018). Low reliability as a threat to the assessment of factorial validity. *Methodology, 14*, 189–197. https://doi.org/10.1027/1614-2241/a000158

Camargo, S., Herrera, A. N., & Traynor, A. (2018). Looking for a consensus in the discussion about the concept of validity – a Delphi study. *Methodology, 14*, 146–155. https://doi.org/10.1027/1614-2241/a000157

Ercikan, K. & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes.* New York, NY: Routledge.

Gadermann, A. M., Chen, M., Emerson, S., & Zumbo, B.D. (2018). Examining validity evidence of self-report measures using Differential Item Functioning: An illustration of three methods. *Methodology, 14*, 165–176. https://doi.org/10.1027/1614-2241/a000156

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55. https://doi.org/10.1080/10705519909540118

Hubley, A. M. (2018, June). *Missed opportunities in testing and assessment: Response processes and test consequences.* Keynote address at the International Congress on Applied Psychology (ICAP), Montréal, Canada, PQ.

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

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

Markus, K. A. (2018). Three conceptual impediments to developing scale theory for formative scales. *Methodology, 14*, 156–164. https://doi.org/10.1027/1614-2241/a000154

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning. Multivariate applications series* (1st ed.). New York, NY: Routledge.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives, 15*, 51–69. https://doi.org/10.1080/15366367.2017.1348108

Newton, P. E. (2012). Clarifying the consensus definition of validity: Commentary. *Measurement: Interdisciplinary Research and Perspectives, 10*, 1–29. https://doi.org/10.1080/15366367.2012.669666

Newton, P. E., & Baird, J. (2016). Editorial: The great validity debate. *Assessment in Education: Principles, Policy & Practice, 23*, 173–177. https://doi.org/10.1080/0969594X.2016.1172871

Rammstedt, B., & Beierlein, C. (2014). Can't we make it any shorter? The limits of personality assessment and way to overcome them. *Journal of Individual Differences, 35*, 212–220. https://doi.org/10.1027/1614-0001/a000141

Rammstedt, B., Beierlein, C., Brähler, E., Eid, M., Hartig, J., Kersting, M., Menold, N., … Weichselgartner, E. (2015). *Quality standards for the development, application, and evaluation of measurement instruments in social science survey research: Prepared and written by the Quality Standards Working Group.* RatSWD Working Papers 245. Retrieved from http://www.ratswd.de/dl/RatSWD_WP_245.pdf

Schultze, M., & Eid, M. (2018). Identifying measurement invariant item sets in cross-cultural settings using an automated item selection procedure. *Methodology, 14*, 177–188. https://doi.org/10.1027/1614-2241/a000155

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

Zumbo, B. D. (2014). What role does, and should, the test standards play outside of the United States of America? *Educational Measurement: Issues and Practice, 33*, 31–33. https://doi.org/10.1111/emip.12052

Zumbo, B. D. & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences. Social Indicators Research Series: Vol. 54*. Cham, Switzerland: Springer.

Zumbo, B. D. & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research. Social Indicators Research Series: Vol. 54.* Cham, Switzerland: Springer.



**Natalja Menold**
GESIS – Leibniz Institute for the Social Sciences, Survey Design and Methodology
PO Box 12 21 55
68072 Mannheim
Gemany
natalja.menold@gesis.org

**Matthias Bluemke**
GESIS – Leibniz Institute for the Social Sciences,
Survey Design and Methodology
PO Box 12 21 55
68072 Mannheim
matthias.bluemke@gesis.org

**Anita M. Hubley**
Department of Educational & Counselling Psychology & Special Education
University of British Columbia
2125 Main Mall
Vancouver, BC V6T 1Z4
Canada
anita.hubley@ubc.ca