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

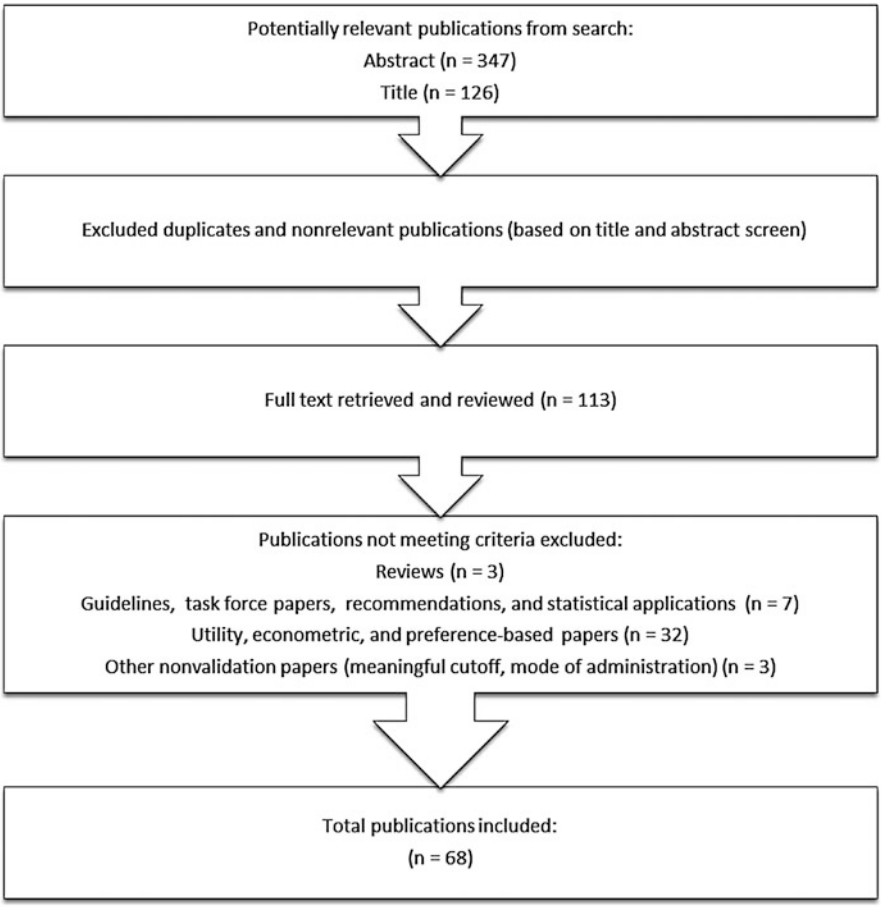

Fig. 15.1 Search process flowchart

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.