Social Indicators Research Series 69

Bruno D. Zumbo
Anita M. Hubley *Editors*

Understanding and Investigating Response Processes in Validation Research

Springer

Social Indicators Research Series

Volume 69

Series Editor
Alex C. Michalos, Faculty of Arts Office, Brandon University, Brandon, Manitoba, Canada

Editors
Ed Diener, University of Illinois, Champaign, USA
Wolfgang Glatzer, J.W. Goethe University, Frankfurt am Main, Germany
Torbjorn Moum, University of Oslo, Norway
Mirjam A.G. Sprangers, University of Amsterdam, The Netherlands
Joachim Vogel, Central Bureau of Statistics, Stockholm, Sweden
Ruut Veenhoven, Erasmus University, Rotterdam, The Netherlands

This 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.

More information about this series at http://www.springer.com/series/6548

Bruno D. Zumbo • Anita M. Hubley
Editors

Understanding and Investigating Response Processes in Validation Research

Springer

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

Anita M. Hubley
Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counselling Psychology, and Special Education (ECPS)
The University of British Columbia
Vancouver, BC, Canada

ISSN 1387-6570 ISSN 2215-0099 (electronic)
Social Indicators Research Series
ISBN 978-3-319-56128-8 ISBN 978-3-319-56129-5 (eBook)
DOI 10.1007/978-3-319-56129-5

Library of Congress Control Number: 2017939937

© Springer International Publishing AG 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Tests and measures are widely used for decision-making, ranking, and policy purposes broadly in the social and behavioral sciences including, more specifically, large-scale testing, assessment, social and economic surveys, and research in psychology, education, health sciences, social and health policy, international and comparative studies, social indicators and quality of life. This is the second book in this series that is wholly focused on validity theory and validation practices. The first book was edited by Zumbo and Chan (2014) and is titled *Validity and Validation in Social, Behavioral, and Health Sciences*. Zumbo and Chan's book is groundbreaking for having focused on the scholarly genre of validation reports and how this genre frames validity theory and validation practices. This second book builds on the themes and findings of the first, with a focus on measurement validity evidence based on response processes.

The *Test Standards* (AERA, APA, & NCME, 2014) presents five sources of validity evidence: content-related, response processes, internal structure, relationships with other variables, and consequences. Zumbo and Chan (2014) showed that response processes validity evidence is poorly understood by researchers and is reported relatively rarely compared to other sources of evidence (e.g., internal structure and relationships to other variables). With an eye toward aiding researchers in providing this type of evidence, this volume presents models of response processes as well as exemplars and methodological issues in gathering response processes evidence. This is the first book to bring together groundbreaking models and methods, including approaches that are novel forms of evidence, such as response shift.

This edited volume is comprised of 19 chapters, including an opening chapter that sets the stage and provides the reader with a description and discussion of response processes validity evidence. The chapters were purposefully chosen to reflect canonical forms of response processes methods as well as a variety of novel research methods and applications. We ordered the chapters in the book alphabetically (by the last name of the first author of the chapter, except, of course, for the opening chapter). In the process of editing the book, we came to the conclusion that any subsections or ordering based on themes and focus were not only artificial but somewhat misleading to the reader – for example, a chapter could be in more than

one subsection. We realize, of course, that grouping and ordering are helpful ways to read and think through the contents of a book. With that in mind, we offer one possible way of organizing the chapters into non-mutually exclusive categories. One could envision five categories:

1. A collection of chapters that provide a description and critical analysis of canonical forms of evidence and methodology (Hubley & Zumbo *opening chapter*; Bruckner & Pellegrino; Leighton et al.; Li et al.; Padilla & Leighton)
2. A collection of chapters that challenge the conceptualization and process of response processes validation (Chen & Zumbo; the two chapters by Launeanu & Hubley; Maddox & Zumbo)
3. A collection of chapters that expand and extend the range of methods used (Chen & Zumbo; Hubley et al.; Li et al.; Padilla & Benitez; Russell & Hubley; Sawatzky et al.; Shear & Roussos; Wu & Zumbo; Zumbo et al.)
4. A collection of chapters that apply response process validation to new research contexts such as business and economics education, writing processes, health psychology, and health surveys/patient-reported outcomes (Bruckner & Pellegrino; Zhang et al.; Zumbo et al.; Beauchamp & McEwan; Sawatzky et al.)
5. A collection of chapters that focus on the statistical models used in response processes validation studies (Chen & Zumbo; Hubley et al.; Li et al., Sawatzky et al.; Wu & Zumbo; Zhang et al.; Zumbo et al.; Zumbo)

Of course, other categorizations of the chapters could be created and may be more useful for readers, but we offer this one as starting point.

Because of its breadth of scope on the topic of response processes as measurement validity evidence, this book is unique in the literature and a high watermark in the history of measurement, testing, and assessment. The chapters clearly have a focus on model building and model testing (be it statistical, cognitive, social psychological, or anthropologic) as central to validation efforts. This focus on validation practices is interesting in and of itself and will influence both future validation studies and theorizing in validity.

We would like to close by acknowledging the impressive body of work that the chapter authors have brought to this volume. We would like to thank Sophie Ma Zhu and Ayumi Sasaki for their assistance with the survey of the studies reporting response processes and with the editing and APA style. In addition, we would like to thank Alex Michalos, the book series editor, as well as Myriam Poort, Esther Otten, and Joseph Daniel from Springer Press.

Vancouver, BC, Canada

Bruno D. Zumbo
Anita M. Hubley

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing* (5th ed.). Washington, DC: American Educational Research Association.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences*. New York, NY: Springer.

Contents

Contributors

Jayanti Banerjee Paragon Testing Enterprises, Inc., Vancouver, BC, Canada

Mark R. Beauchamp Psychology of Exercise, Health, and Physical Activity Laboratory, School of Kinesiology, War Memorial Gym, The University of British Columbia, Vancouver, BC, Canada

Isabel Benítez Universidad Loyola Andalucía, Sevilla, Spain

Ronak Brahmbhatt Ted Rogers School of Management, Ryerson University, Toronto, ON, Canada

School of Nursing, Trinity Western University, Langley, BC, Canada

Sebastian Brückner Department of Business and Economics Education, Johannes Gutenberg-University Mainz, Mainz, Rheinland-Pfalz, Germany

Eric K.H. Chan School of Nursing, Trinity Western University, Langley, Canada

Measurement, Evaluation, and Research Methodology (MERM) Program, The University of British Columbia, Vancouver, Canada

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

Paul Deane MS T03, Educational Testing Service, Princeton, NJ, USA

Barry Forer The Human Early Learning Partnership, The University of British Columbia, Vancouver, BC, Canada

Qi Guo Center for Research in Applied Measurement and Evaluation (CRAME), Department of Educational Psychology, Faculty of Education, University of Alberta, Edmonton, AB, Canada

Anita M. Hubley Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counselling Psychology, and Special Education (ECPS), The University of British Columbia, Vancouver, BC, Canada

Mihaela Launeanu MA Counselling Psychology Program, Trinity Western University, Langley, BC, Canada

Jacqueline P. Leighton Center for Research in Applied Measurement and Evaluation (CRAME), Department of Educational Psychology, Faculty of Education, University of Alberta, Edmonton, AB, Canada

Chen Li MS T03, Educational Testing Service, Princeton, NJ, USA

Zhi Li Paragon Testing Enterprises, Inc., Vancouver, BC, Canada

Yan Liu Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counselling Psychology, and Special Education (ECPS), The University of British Columbia, Vancouver, BC, Canada

Lisa M. Lix Department of Community Health Sciences, Rady Faculty of Health Sciences, University of Manitoba College of Medicine, Winnipeg, MB, Canada

Bryan Maddox School of International Development, University of East Anglia, Norwich, UK

Laboratory of International Assessment Studies, Norwich, UK

Desmond McEwan Psychology of Exercise, Health, and Physical Activity Laboratory, School of Kinesiology, War Memorial Gym, The University of British Columbia, Vancouver, BC, Canada

José-Luis Padilla University of Granada, Granada, Spain

James W. Pellegrino University of Illinois at Chicago, Learning Science Research Institute, Chicago, IL, USA

Louis A. Roussos Measured Progress, Dover, NH, USA

Lara B. Russell Centre for Health Evaluation and Outcome Sciences, Providence Health Care, St. Paul's Hospital, Vancouver, BC, Canada

Tolulope T. Sajobi Department of Community Health Sciences, University of Calgary, Calgary, AB, Canada

Richard Sawatzky School of Nursing, Trinity Western University, Langley, BC, Canada

Centre for Health Evaluation and Outcome Sciences, Providence Health Care, Vancouver, BC, Canada

Benjamin R. Shear School of Education, University of Colorado Boulder, Boulder, CO, USA

Wei Tang Center for Research in Applied Measurement and Evaluation (CRAME), Department of Educational Psychology, Faculty of Education, University of Alberta, Edmonton, AB, Canada

Amery D. Wu Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counselling Psychology, and Special Education (ECPS), The University of British Columbia, Vancouver, BC, Canada

Mo Zhang MS T03, Educational Testing Service, Princeton, NJ, USA

Danjie Zou Measurement, Evaluation, and Research Methodology (MERM) Program, Department of Educational and Counselling Psychology, and Special Education (ECPS), The University of British Columbia, Vancouver, BC, Canada

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

Chapter 1
Response Processes in the Context of Validity: Setting the Stage

Anita M. Hubley and Bruno D. Zumbo

Opening Remarks

Tests and measures are widely used for decision-making, ranking, and policy purposes in the social and behavioral sciences using large-scale testing, regularly administered tests of a population over time, assessment of individuals, as well as social and economic surveys. These sorts of studies are conducted in disciplines such as psychology, education, health sciences, social and health policy, international and comparative studies, social indicators and quality of life, to name but a few. Zumbo and Chan (2014) showed that approximately 1000 studies are published each year examining the validity of inferences made from tests and measures in the social, behavioral, and health sciences. The *Standards for Educational and Psychological Testing*[1] (AERA, APA, & NCME, 2014) provides a description and a set of standards for validation research. Although the *Standards* (AERA et al., 2014) were developed in the United States and with test development and test use in that country in mind, they have impact worldwide (Zumbo, 2014). The *Standards* present five sources of evidence for validity: test content, response processes, internal structure, relations to other variables, and consequences of testing. Zumbo and Chan, and the various contributors to their volume, showed that many studies focus on internal structure and relations with other variables sources of evidence, which have a long history in validation research, are known methodologies, and have numerous exemplars in the literature. Far less is understood by test users and researchers conducting validation work about how to think about and apply new and

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

A.M. Hubley (✉) • B.D. Zumbo
Measurement, Evaluation, and Research Methodology (MERM) Program,
Department of Educational and Counselling Psychology, and Special Education (ECPS),
The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
e-mail: anita.hubley@ubc.ca; bruno.zumbo@ubc.ca

© Springer International Publishing AG 2017

B.D. Zumbo, A.M. Hubley (eds.), *Understanding and Investigating Response Processes in Validation Research*, Social Indicators Research Series 69,
DOI 10.1007/978-3-319-56129-5_1

emerging sources of validity evidence. As we will discuss more fully below, evidence based on response processes is both important and most illuminating in building a strong body of evidence for the validity of the inferences from our tests and measures.

The remainder of this chapter is organized into four sections. The first section addresses the all-important, and largely ignored, question of what are response processes. It is remarkable that discussions of, and research on, response processes have gone on for so many years without a well-accepted definition expressed in the literature. The second section takes an 'over the shoulder look' back at some key moments in the history of response processes. It is advisable, if not illuminating, to set a course forward by at least glancing at where we have been. The third section reports on the prevalence of the reporting of evidence based on response processes in the published research literature. And the final section sets a course for the future by asking the question, where do we go next?

What Are Response Processes?

Response processes are one of five sources of validity evidence described in the 1999 and 2014 *Standards* (AERA, APA, & NCME, 1999; AERA et al., 2014). Unlike the 1999 *Standards*, the 2014 *Standards*, however, explicitly references the "*cognitive* processes engaged in by test takers" [italics added] (AERA et al., 2014, p. 15). Both *Standards* suggest that "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 the performance or response actually engaged in by test takers" (e.g., AERA et al., 2014, p. 15). Surprisingly though, the *Standards* do not provide a clear conceptual or operational definition of response processes; rather, they focus on the techniques and methods one may use to obtain validity evidence using response processes as a source.

Clearly, the most attention in response processes research has been paid to cognitive models of responding. This has been evident in the longstanding research program of Susan Embretson (e.g., Embretson, 1983, 1984, 1993; Embretson, Schneider, & Roth, 1986), but also influenced by research by James Pellegrino (e.g., Pellegrino, Baxter, & Glaser, 1999; Pellegrino, DiBello, & Goldman, 2016; Pellegrino & Glaser, 1979), and Robert Mislevy (e.g., Mislevy, 2009; Mislevy, Steinberg, & Almond, 2002). Brückner and Pellegrino (2016) point out response processes may consist of multiple mental operations (which are measurable and neurobiologically based) and phases.

We argue, however, that one may think broadly of response processes as the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed test score variation. This definition expands response processes beyond the cognitive realm to include emotions, motivations, and behaviors. Inclusion of affect and

motives allows us to take into account how these may impact the different respondents' interactions with the item(s), test, and testing situation. Our definition also requires one to go beyond the surface content of the actions, thoughts, or emotions expressed by, or observed in, respondents to identify the mechanisms that underlie this content. Finally, we encourage researchers and theorists to develop contextualized and dynamic frameworks that take into account the situational, cultural, or ecological aspects of testing when exploring evidence based on response processes.

In considering what response processes are, it is also important to point out what they are not. In the medical education field, Downing (2003) is a commonly cited source on validity evidence. Downing defines response process as "evidence of data integrity such that all sources of error associated with the test administration are controlled or eliminated to the maximum extent possible" (p. 834), including, for example, quality control of data, documentation of practice materials, appropriateness of methods used to combine scores into a composite score, and explanations and interpretive materials provided when reporting scores. Although Downing claims to rely on the *Standards* (AERA et al., 1999) in his presentation, it is not clear how he came to interpret response processes the way he has as this, in no way, resembles how response processes are described in the *Standards* (AERA et al., 1999, 2014). What Downing is talking about is really technical and procedural quality; this may influence reliability and validity but it is not response processes and we strongly discourage researchers and test users from applying his operational definition because it conflates too many different measurement ideas that are not, themselves, validity. Still, Downing's interpretation of response processes has been cited in other articles describing the kinds of evidence that can be used to support different sources of validity evidence (e.g., Cook & Beckman 2006; Cook, Zendejas, Hamstra, Hatala, & Brydges, 2014).

It is also important not to confuse a definition of response processes with the techniques and methods used to obtain such evidence. Because of the focus on cognitive processes, using cognitive interviewing, think aloud protocols, and Cognitive Aspects of Survey Methodology (CASM; Tourangeau, 1984) have seemed a natural way to capture this, and response processes research has become intrinsically intertwined with these methods of late. There are other techniques and methods for obtaining validity evidence based on response processes as described by the *Standards* (AERA et al., 2014), Messick (1989b); Padilla and Benítez (2014), and many of the chapter authors in this volume. Some of these other methods include: response times; eye tracking methods; keeping records that track the development of a response; analyzing the relationship among components of a test or task, or between test scores and other variables, that address inferences about processes; paradata (e.g., mouse clicks, accessing definitions or explanations, changing responses); anthropological data (e.g., stance, position, glances, gestures); and statistical, psychometric, or computational response process models. However, the examination of response processes is not limited to the respondents. The *Standards* (AERA et al., 2014) also note that, if a measure relies on observers, scorers, or

judges to evaluate respondents, then the psychological or cognitive processes used by these observers, scorers, or judges should be examined to determine if they are consistent with the intended interpretation of scores. This may include the use of cognitive interviewing and think-aloud protocols, documenting or recording responses to items, recording the time needed to complete the task of the observers, scorers, or judges, and follow-up questionnaires or interviews.

A final comment is needed about connections between response processes and content validation. Hubley, Zhu, Sasaki, and Gadermann (2014) pointed out that some researchers seem to blur evidence that is based on response processes with evidence based on test content. Whether one might view response processes evidence as forming an independent source of validity evidence or an element of content validation depends on how one views the realm of content validation (see, for example, Padilla & Benítez, 2014). Much of this confusion may stem from Messick's (1989a, 1995) work in which he has been somewhat unclear on the role of response processes; that is, he sometimes treats response processes as evidence that elevates test content in contributing to construct validity and sometimes as separate evidence that is linked to or informs test content (e.g., see Messick, 1995 and his discussions of representativeness as a core concept that links his content and substantive aspects of construct validity).

Key Moments and Players in the History of Response Processes

Roger Lennon

Most descriptions of response processes as validation evidence attribute the concept to Samuel Messick, but the concept of response processes as validation evidence has been around for some time. Lennon (1956) incorporated response processes under content validation, arguing that "appraisal of content validity must take into account not only the *content* of the questions but also the *process* presumably employed by the subject in arriving at his response" (p. 296). Lennon's point was that content validity is about the responses, rather than the items, because the responses reflect the respondent's behaviours.[2] Thus, if different respondents respond using different processes, then content validity may differ among those respondents despite the items being the same.

[2] Messick (1989a, 1990) would agree with this view but noted that the dominant view of content validation focuses on expert judgments about test content representativeness and relevance. It is because the dominant view of content validity does not address response consistencies and test scores that Messick (1989b) argued that "so-called content validity does not qualify as validity at all" (p. 7).

Susan Embretson

By far, the most extensive research program on response processes as evidence for validity, or alternatively that contributes to the description and understanding of test performance, has been conducted by Susan Embretson (e.g., Embretson, 1983, 1984, 1993; Embretson & Schneider, 1989; Embretson et al., 1986; Whitely, 1977). Much of Embretson's work has sought to clarify the validity of inferences made from intelligence, cognitive, aptitude, or neuropsychological tests by treating test items as information-processing tasks. Her research program was clearly impacted by not only cognitive psychology, information processing approaches, and cognitive component analysis, but also by experimental psychology and psychometrics. She generously gives a nod to Messick's early (1972) claim that there is a need in the psychometric field to develop models of psychological processes that underlie test performance (Whitely, 1977).

Embretson (1983) proposed that construct validity is comprised of two aspects: (a) construct representation, and (b) nomothetic span. Construct representation has to do with identifying theoretical mechanisms (e.g., processes, strategies, knowledge stores, metacomponents) that underlie test items or task performance whereas nomothetic span has to do with the network of relationships between the test score(s) and other variables. In the parlance of the *Standards* (AERA et al., 1999, 2014), one might think of construct representation as falling under the response processes source of evidence and nomothetic span as falling under the relations to other variables source of evidence. Embretson (1983) saw construct representation as being concerned with the meaning of test scores whereas nomothetic span has to do with the significance of test scores. Furthermore, she and her colleagues argued that the theoretical mechanisms can be examined using methods of task decomposition from information processing (Embretson et al., 1986).

To examine construct representation, Embretson and her colleagues (Embretson, 1984; Embretson & Yang, 2013; Whitely, 1980) developed and implemented elaborate noncompensatory and compensatory multicomponent latent trait psychometric models for cognitive diagnosis that can be used to test hypotheses about attributes and skills thought theoretically to underlie response processes (e.g., difficulty).

There are further exemplars of the marriage of cognitive psychology and psychometric theory in Embretson's more recent work with colleagues that extends the use of response processes evidence (e.g., Gorin & Embretson, 2006; Ivie & Embretson, 2010). In the former, they introduce a new technology called algorithmic item generation in which items are systematically created based on specific combinations of features that underlie the processing required to correctly solve a problem. In both papers, data are gathered and statistical models are fit to examine the contribution of item characteristics to the difficulty of the item with an eye toward possible aspects of item design useful for future developments in item generation.

Samuel Messick

Messick (1995) identified six aspects of construct validity that function as general validity standards for educational and psychological measurement. Messick (1995) incorporated response processes under his substantive aspect of construct validity, which he argued "refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks" (p. 745). Messick (1995) further argued that we need to move beyond the use of expert judgments of content to gather evidence that the processes we claim to have sampled are actually engaged by respondents when responding to items or tasks.

Importantly, Messick (1995) described construct validity as comprising "the evidence and rationales supporting the trustworthiness of score interpretation in terms of *explanatory concepts* that account for both test performance and score relationships with other variables" [italics added] (p. 743). He noted that, historically, most attention has been placed on evidence involving essentially internal structure, convergent and discriminant coefficients, and test-criterion relationships, but that evidence of expected differences in performance over time, across settings or groups, and as a result of experimental manipulation would be more illuminating. He then pointed out that "possibly most illuminating of all, however, are direct probes and modeling of the processes underlying test responses…At the simplest level, this might involve querying respondents about their solution processes or asking them to think aloud while responding to exercises during field trials" (p. 743). Messick (1989a) further pointed out that similarities and differences in response processes can be examined across groups or contexts as well as over time to provide evidence for the generalizability of test score interpretation and use. Messick (1995) also made it clear that no matter what evidence is used to contribute to understanding score meaning, "the contribution becomes stronger if the degree of fit of the information with the theoretical rationale underlying score interpretation is explicitly evaluated" (p. 743). These descriptions of response processes as a source of validity evidence highlight its important role in construct validation, the strength of the evidence that it can provide, guidance that verbal reports (e.g., cognitive interviewing, think aloud protocols) are just a starting point with further evidence needed, and the important role of examining fit between what is theoretically expected and what is found when respondents interact with items and tasks of given constructs.

Standards for Educational and Psychological Testing and Other Guidelines

The first time that response processes appear in the *Standards* is in the 1985 edition (APA, AERA, & NCME, 1985), but they are only included as evidence of construct validity. Response processes first appeared as one of five sources of validity

evidence in the 1999 *Standards* (AERA et al., 1999). Those five sources remained unchanged in the 2014 *Standards*, as does most of the information on response processes (AERA et al., 2014). It is unclear why, or what was going on in discussions about validity and validation within or outside of the joint committee on the *Standards*, that response processes were elevated from a form of evidence in the 1985 edition of the *Standards* to one of the five main sources of evidence in the 1999 *Standards*.

Chan (2014), in his review of standards and guidelines for validation practices, found only two other groups that subsequently and explicitly included response processes as evidence; that is, the Society for Industrial and Organizational Psychology's (SIOP) *Principles for the Validation and Use of Personnel Selection Procedures*, and the Buros Center for Testing's *Mental Measurements Yearbook*.

Prevalence of Validity Evidence Based on Response Processes

Only recently have validation syntheses started to document the prevalence of validity evidence based on response processes. Beckman, Cook, and Mandrekar (2005) conducted a search of various databases, MEDLINE, EMBASE, PsycINFO, ERIC, and the Social Science Citation/Science Citation indices for psychometric articles on assessments of clinical teaching published between 1966 and mid-2004. Of the 22 relevant studies, only two provided evidence of response processes. Cizek, Rosenberg, and Koon (2008) reviewed 283 tests from the 16th *Mental Measurements Yearbook* produced by the Buros Institute of Mental Measurements. They found that evidence based on response processes was mentioned in only 1.8% of the cases. Villalobos Coronel (2015) examined 30 psychometric studies from 27 articles conducted on the Rosenberg Self-Esteem Scale from 1989 to 2015; validity evidence based on response processes was reported in only 1 (3.3%) study.

Recently, Zumbo and Chan (2014) edited a volume of 15 research syntheses of validity evidence reported in a variety of research areas. Chapters in the book tended to focus on syntheses of evidence from specific journals or from specific measures. It is abundantly evident from the various chapters that response processes evidence is sorely neglected (see also Lyons-Thomas, Liu, & Zumbo, 2014). Many syntheses found no evidence of response processes evidence being reported (e.g., Chan, Munro, et al., 2014; Chan, Zumbo, Chen, et al., 2014; Chan, Zumbo, Zhang, et al., 2014; Collie & Zumbo, 2014; Cox & Owen, 2014; Gunnell, Wilson, et al., 2014; Hubley et al., 2014). Slightly more chapters found some evidence of response processes evidence being reported, but it was very limited and tended to only include 1–3 of all of the studies examined in each case (e.g., Ark, Ark, & Zumbo, 2014; Chan, Zumbo, Darmawanti, & Mulyana, 2014; Chinni & Hubley, 2014; Gunnell, Schellenberg, et al., 2014; Hubley et al., 2014; McBride, Wiens, McDonald, Cox, & Chan, 2014; Sandilands & Zumbo, 2014; Shear & Zumbo, 2014; Zumbo et al., 2014).

There has been an influx of research incorporating evidence based on response processes in the last 5 years. Much of this work has emerged in the medical education

field. Because this work tends to cite Downing (2003) as a source, some concern must be expressed about whether many of these studies actually provide response processes based evidence as defined here and commonly accepted in the validity field. Thus, response processes evidence that relies solely on technical and procedural quality information, such as inter-rater reliability estimates, documentation of scoring, or justification for use of a composite score, may inflate, and thus incorrectly reflect, the prevalence of validity evidence based on response processes.

Still, it is clear from this brief overview of recent research that very few studies have attended to validity evidence that stems from response processes. As noted by Hubley et al. (2014), one reason why relatively few studies have been conducted that report validity evidence based on response processes is that, relative to the other sources of validity evidence, there is less clear and accepted practice about how to design such studies or how to report them. Moreover, it is difficult to locate such evidence in the literature, especially if easily identifiable or clear keywords (e.g., response processes, validity, validation) are not associated with these studies or materials.

Where Do We Go Next?

It is clearly time that greater attention be paid to theorizing about, and gathering validity evidence based on, response processes. To date, a lot of work in response processes has been descriptive. What is missing is an understanding of why people respond the way that they do; that is, research in response processes needs to become more explanation-based. Identifying and understanding the mechanisms underlying how different respondents interact with, and respond to, test items and tasks is essential to understanding score meaning and test score variation. This research needs to not only take into account what happens narrowly in the generative space and time between when the test taker sees the item and the response is completed but also the broader context (i.e., purpose of testing, setting, culture) that influences the respondent, the test, and the test interpretation.

This groundbreaking volume, *Understanding and Investigating Response Processes in Validation Research,* addresses an urgent need across multiple disciplines to broaden our understanding and use of response processes as a source of evidence for the validity of inferences made from test scores. This volume presents conceptual models of response processes, methodological issues that arise in gathering response processes evidence, as well as applications and exemplars for providing response processes evidence in validation work. The collection of chapters shows the reader how to conceptualize response processes while encouraging the reader to reflect critically on validity evidence. Novel forms of response processes evidence are introduced and examples are provided for how to design and report response processes evidence. A key feature of the collection of chapters is that it counters the nature of measurement research as silos in sub-disciplines and shows how response processes evidence is relevant and applicable to a wide range of disciplines in the social, behavioral, and health sciences.

This volume reflects a paradigmatic shift in validation research and response processes validation, in particular. There are several key messages that will serve as points of interest as we venture forward in response processes validation research. First, treating the field of measurement, testing, and assessment as distinct sub-disciplinary silos is not productive. Acknowledging that the different sub-disciplines (e.g., language testing, educational testing, psychological assessment, health measurement, patient-reported outcomes, and medical education) have uniquenesses governed by their particular domains and applications, it is important to note that they have far more in common. Most importantly, in using the common language of validity and validation, we have the opportunity to learn from the measurement challenges that arise in of each of these sub-disciplinary contexts and can build on those in the advances we make in validity theory and practice. In this light, we agree with Zumbo (2014) that the globalization of the *Standards* (AERA et al., 2014) allows them to play a key role in the measurement, test, and assessment community worldwide and should serve both as a common source of terminology and as a touch-stone as we move forward.

A second key point of interest as we move forward is that the expanding notions of response processes offered in this volume challenge the boundaries of our current conceptualizations of responses processes and expand the evidential basis and methodology beyond the canonical methods of mental probes afforded by think aloud protocols and cognitive interviewing. In the end, it becomes apparent that not all response processes evidence need be based solely on individuals or be purely mentalistic. The key feature is adopting a scientific mindset and developing and testing explanatory models of response processes for test validation purposes. This necessitates an appreciation for what models are and how they serve (or might serve) in assembling evidence for response processes. Moreover, given the wide range of disciplines in which assessments, tests, and measures are used, the set of possible models and modeling practices needs to be inclusive of: (i) cognitive models, (ii) ecological, contextualized, and environmental perspectives to modeling, (iii) novel disciplinary contributions such as anthropologic models that focus on, for example, stance or gesture, (iv) affective and motivational models, (v) elaborated statistical or mathematical models that take into account the complex settings of real-life test-taking, and (vi) a re-casting of our psychometric models (such as item response theory) back to their early focus on describing the response process. In short, the use of explanatory models helps us both (a) view items and assessment tasks as windows into the minds of test respondents, and (b) understand and describe the enabling conditions for item responses.

References

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

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.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education [APA, AERA, & NCME]. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ark, T. K., Ark, N., & Zumbo, B. D. (2014). Validation practices of the Objective Structured Clinical Examination (OSCE). In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 267–288). New York, NY: Springer.

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. doi:10.1111/j.1525-1497.2005.0258.x.

Brückner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment of higher education students' competency in business and economics. *Journal of Educational Measurement, 53*, 293–312. doi:10.1111/jedm.12113.

Chan, E. K. H. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 9–24). New York, NY: Springer.

Chan, E. K. H., Munro, D. W., Huang, A. H. S., Zumbo, B. D., Vojdanijahromi, R., & Ark, N. (2014). Validation practices in counseling: Major journals, mattering instruments, and the Kuder Occupational Interest Survey (KOIS). In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 67–87). New York, NY: Springer.

Chan, E. K. H., Zumbo, B. D., Chen, M. Y., Zhang, W., Darmawanti, I., & Mulyana, O. P. (2014). Reporting of measurement validity in articles published in *Quality of Life Research*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 217–228). New York, NY: Springer.

Chan, E. K. H., Zumbo, B. D., Darmawanti, I., & Mulyana, O. P. (2014). Reporting of validity evidence in the field of health care: A focus on papers published in *Value in Health*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 257–265). New York, NY: Springer.

Chan, E. K. H., Zumbo, B. D., Zhang, W., Chen, M. Y., Darmawanti, I., & Mulyana, O. P. (2014). Medical Outcomes Study Short Form-36 (SF-36) and the World Health Organization Quality of Life (WHOQoL) assessment: Reporting of psychometric validity evidence. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 243–255). New York, NY: Springer.

Chinni, M., & Hubley, A. M. (2014). The Satisfaction with Life Scale (SWLS): A review of reported validation practice. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 35–66). New York, NY: Springer.

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.

Collie, R. J., & Zumbo, B. D. (2014). Validity evidence in the *Journal of Educational Psychology*: Documenting current practice and a comparison with earlier practice. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 113–135). New York, NY: Springer.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine, 119*, 166.e7–166.e16. http://doi.org/10.1016/j.amjmed.2005.10.036.

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education, 19*, 233–250. doi:10.1007/s10459-013-9458-4.

Cox, D. W., & Owen, J. J. (2014). Validity evidence for a perceived social support measure in a population health context. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 229–241). New York, NY: Springer.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*, 830–837. doi:10.1046/j.1365-2923.2003.01594.x.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175–186. doi:10.1007/BF02294171.

Embretson, S. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Embretson, S., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement, 23*, 13–32. doi:10.1111/j.1745-3984.1986.tb00231.x.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Embretson, S. E., & Schneider, L. M. (1989). Cognitive component models for psychometric analogies: Conceptually driven versus interactive process models. *Learning and Individual Differences, 1*, 155–178. doi:10.1016/1041-6080(89)90001-0.

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*, 14–36. doi:10.1007/s11336-012-9296-y.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*, 394–411. doi:10.1177/0146621606288554.

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 & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 137–156). New York, NY: Springer.

Gunnell, K. E., Wilson, P. M., Zumbo, B. D., Crocker, P. R. E., Mack, D. E., & Schellenberg, B. J. I. (2014). Validity theory and validity evidence for scores derived from the Behavioural Regulation in Exercise Questionnaire. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 175–191). New York, NY: Springer.

Hubley, A. M., Zhu, M., Sasaki, A., & Gadermann, A. (2014). A synthesis of validation practices in the journals *Psychological Assessment* and *European Journal of Psychological Assessment*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 193–213). New York, NY: Springer.

Ivie, J. L., & Embretson, S. E. (2010). Cognitive process modeling of spatial ability: The assembling objects task. *Intelligence, 38*, 324–335. doi:10.1016/j.intell.2010.02.002.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*, 294–304.

Lyons-Thomas, J., Liu, Y., & Zumbo, B. D. (2014). Validation practices in the social, behavioral, and health sciences: A synthesis of syntheses. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (p. 313319). New York, NY: Springer.

McBride, H. L., Wiens, R. M., McDonald, M. J., Cox, D. W., & Chan, E. K. H. (2014). The Edinburgh Postnatal Depression Scale (EPDS): A review of the reported validity evidence. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 157–174). New York, NY: Springer.

Messick, S. (1972). Beyond structure: In search of functional models of psychological process. *Psychometrika, 37*, 357–375.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5–11.

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

Messick, S. (1990). *Validity of test interpretation and use*. Princeton, NJ: Educational Testing Service.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 83–108). Charlotte, NC: Information Age.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the role of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 97–128). Mahwah, NJ: Lawrence Erlbaum.

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*, 136–144.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24*, 307–353.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51,* 59–81. doi:10.1080/00461520.2016.1145550.

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence, 3*, 187–215.

Sandilands, D., & Zumbo, B. D. (2014). (Mis)alignment of medical education validation research with contemporary validity theory: The Mini-CEX as an example. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 289–310). New York, NY: Springer.

Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in *Educational and Psychological Measurement*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 91–111). New York, NY: Springer.

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academy Press.

Villalobos Coronel, M. (2015). *Synthesis of reliability and validation practices used with the Rosenberg self-esteem scale*. Master's thesis, University of British Columbia. Retrieved from https://open.library.ubc.ca/cIRcle/collections/24/items/1.0165784

Whitely (Embretson), S. E. (1977). Information-processing on intelligence test items: Some response components.*AppliedPsychologicalMeasurement,1*,465–476.doi:10.1177/014662167700100402.

Whitely (Embretson), S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479–494.

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.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences*. New York, NY: Springer.

Zumbo, B. D., Chan, E. K. H., Chen, M. Y., Zhang, W., Darmawanti, I., & Mulyana, O. P. (2014). Reporting of measurement validity in articles published in *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 27–34). New York, NY: Springer.

Chapter 2
Response Processes and Measurement Validity in Health Psychology

Mark R. Beauchamp and Desmond McEwan

Within the field of health psychology, researchers and practitioners are broadly concerned with the array of psychological, environmental, and behavioural factors that contribute to the presence or absence of health (i.e., illness) across diverse life contexts, as well as various means of intervention that can be used to enhance health in these different settings. In order to achieve these broad and laudable goals it is essential that researchers and practitioners have at their disposal measurement devices that are able to provide reliable and valid information about the target variable being assessed. A wide range of measurement approaches that are often used include observations of behavior (e.g., patient compliance checklists), healthcare records (morbidity, mortality), physiological assessments (blood pressure, body composition), psychophysiological assessments (functional magnetic resonance imaging), as well as questionnaires that assess various psychological processes (Johnston, French, Bonetti & Johnston, 2004). It is with respect to this latter research methodology that represents the focus of examination in this chapter and, in particular, the methodological procedures that are used to maximize the reliability and validity of inferences derived from responses to psychological assessments.

Broadly considered, validity is concerned with "an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores" (Messick, 1995, p. 741). In crude terms, if measures related to a given (psychological, behavioral, or environmental) variable display solid evidence of validity (is it measuring what we believe that it measures?), one can make inferences about the nature of that variable, how it relates to other constructs, and potentially how that variable can be changed or enhanced through intervention. Of course, the

M.R. Beauchamp (✉) • D. McEwan
Psychology of Exercise, Health, and Physical Activity Laboratory, School of Kinesiology, War Memorial Gym, The University of British Columbia,
122 – 6081 University Blvd., Vancouver V6T 1Z1, BC, Canada
e-mail: mark.beauchamp@ubc.ca; desi.mcewan@ubc.ca

© Springer International Publishing AG 2017

B.D. Zumbo, A.M. Hubley (eds.), *Understanding and Investigating Response Processes in Validation Research*, Social Indicators Research Series 69,
DOI 10.1007/978-3-319-56129-5_2

corollary is, if a given measure displays poor validity, at best we are hindered from fully understanding that construct, and perhaps more damagingly, researchers and practitioners can make erroneous conclusions that lead them to intervene in sub-optimal or problematic ways. In short, measurement validity is critical to the field of health psychology. In this chapter, we examine the importance of *response processes* within a broader/unified validity theory framework (cf. Messick, 1995), and explain how (a) different methodological procedures can be used enhance the validity of measures derived from health psychology assessments (in particular, questionnaires), and (b) a failure to consider and operationalize these methodological processes can potentially be problematic.

Messick's (1995) Unified Validity Theory Framework

Within the field of health psychology, and indeed across other fields of psychology, the use of the term 'validity' has been used in somewhat inconsistent ways. While some have used the term in relation to the validity of instruments or questionnaires, we take the view presented by Messick (1989, 1995) and others (e.g., Smith, 2005) that validity is not a property of a given instrument or questionnaire; rather, it is a property of test scores (i.e., participants' responses) that derive *from* that instrument or questionnaire. Thus, it is the inferences and interpretations made from those responses that are subject to validation (Hubley & Zumbo, 1996; Messick 1995). At the core of Messick's unified view of validity lies *construct validity* which involves "the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables" (Messick, 1995, p. 743). From this perspective, construct validity is concerned with appraising multiple sources of *validity evidence* that include 'content', 'substantive', 'structural', 'generalizability', 'external', and 'consequential' considerations (cf. Messick, 1995).

The first step in developing any questionnaire, or indeed any other assessment procedure (e.g., observational assessment protocol), is to ensure that the questionnaire, and items subsumed within it, directly and accurately reflect the construct (or concept) under investigation. Specifically, the *content* aspect of validity is concerned with *content relevance* and *representativeness,* whereby questionnaire items should be fully representative of, and directly align with, the content of the construct being studied, and no other (i.e., reflecting different, incongruent or misaligned concepts). A critical first step in this process (and before any items are constructed) is to fully articulate the conceptual bases and theoretical framework that is being used to study the very *nature* of the construct under investigation. This might involve articulating the extent to which the construct is conceptually different from other (similar) variables and distinct from conceptual antecedents and consequences, to ensure those predictor and criterion variables do not become conflated with the construct under study. This conceptual framing might also involve a clear explanation of potential boundary conditions (i.e., moderators) and mechanistic

processes (i.e., mediators) that are subsumed within the overall theory. Indeed, as several prominent scholars such as Clark and Watson (1995), Meehl (1990), and Smith (2005) have noted, it is critical that researchers first provide a clear and meaningful explanation of theory, including an "articulation of how the theory of the construct is translated into informative hypotheses" (Smith, 2005, p. 399). Of course 'theories' can be derived through different means; however, without an articulated theory, there is no construct validity (Cronbach & Meehl, 1955).

With theory guiding the subsequent development of items to reflect the target construct, two key steps can be followed to enhance the content aspect of validity. The first is to involve members of the target population in the development and refinement of specific items to ensure that those questionnaire items are both fully representative and relevant to the world views of those persons (Beauchamp et al., 2010; Vogt, King, & King, 2004). The second is to ensure that (arm's-length) experts are involved in critically appraising the extent to which any preliminary pool of items aligns with the theoretical frames underpinning the focal measure, and to further ensure that items are theoretically grounded, insofar as they are fully relevant to, and representative of, the focal construct (Beauchamp, Bray, Eys & Carron, 2002; Messick, 1995).

The *substantive* aspect of validity is concerned with accruing empirical evidence that participants' responses (to questionnaire items) align with what is purported to be measured within a given item, questionnaire, or assessment protocol. For instance, when participants respond to items subsumed within a questionnaire, do their response processes directly correspond with what is contended to be queried within that questionnaire? As an example, recent work within the field of health psychology has challenged whether items that are typically designed, and used, to assess self-efficacy (i.e., beliefs about personal capability) unintentionally assess intention (i.e., motivation) and not the target construct, namely self-efficacy (Williams & Rhodes, 2016). This issue, and ensuing debate, is described in detail in the following section. However, at a very basic level, if respondents interpret questionnaire items in a manner that is different from that intended by the instrument developer (and the over-arching theory), this has non-trivial implications for not only understanding the nature of the focal construct (and how it might relate to other variables), but also has substantive implications for intervention as well as (health, education, and social) policy. There are several methodological strategies available to instrument developers to enhance the substantive aspects of construct validity (cf. Messick, 1995), that include the use of cognitive interviewing to ascertain what respondents are actually thinking 'in situ' while completing responses to questionnaires (Oremus, Cosby, & Wolfson, 2005; Willis, 2005), the use of implicit measures (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009), as well as behavioural measures (Mayer, Salovey, Caruso, & Sitarenios, 2003). Attending to the substantive aspects of validity and determining that participants' responses to assessment align with what is purported to be assessed, ensures that a strong foundation is provided before any subsequent psychometric and applied research is conducted. Indeed, as we will illustrate in the next section of this chapter, failing to seriously consider the substantive aspects of validity can undermine any efforts to

ascertain the 'structural', 'generalizability', and 'external' aspects of validity, resulting in non-trivial consequences for theory/hypothesis testing and indeed intervention, in what Messick (1995) and others (cf. Hubley & Zumbo, 2011) have referred to as 'consequential' validity concerns.

The structural aspects of validity are concerned with evidence that is based on the internal structure of measures derived from a given instrument. This might be ascertained through examination of model-data fit through factor analysis, item loadings, inter-factor correlations, and so forth (Hu & Bentler, 1999). The generalizability aspect of validity is concerned with the extent to which inferences derived from test scores can in fact be generalized to other populations and contexts. For example, if extensive validity evidence is derived in support of a given questionnaire among a sample of working-age adults to what extent might those findings, and inferences derived from those findings, be applicable to other groups such as teenagers or older adults? The external aspect of construct validity is concerned with examining evidence based on the relations between measures of the focal construct and measures derived in relation to other relevant variables. With this in mind, the external aspect of validity is concerned with both applied utility and criterion relevance. Specifically, external aspects of validity are concerned with examining the extent to which measures derived in relation to a focal construct predict and explain variance in theoretically relevant variables and/or contribute to discriminant utility by displaying divergence with measures derived from theoretically unrelated variables. Finally, the consequential aspects of validity are concerned with examining the various and broad reaching (often unintended) implications that might be derived from use of a particular test.

Across diverse spheres of human functioning, there are numerous examples of (unintended) consequences that have arisen from the use of various assessment procedures. As one example, as a result of the No Child Left Behind Act of 2001 in the United States, all states were required to administered standardized tests in reading and mathematics in Grades 3 and 8, on the premise that such tests would help to raise standards. As Schwartz (2015) recently noted "supporters of this approach were not out to undermine the engagement, creativity, and energy of good teachers." (p. 45). What resulted however, was not only a narrowed curricula whereby teachers 'taught-to-the-test' (and forgoing teaching and learning that fell outside of the curricula) but, with student performances on these tests tied to teacher salaries/bonuses and even the fate of some schools, instances arose of (some) teachers cheating by changing students' answers to exam questions (Schwartz, 2015). In the health field, an example of consequences associated with test administration comes from the recent emergence of direct-to-consumer (DTC) genetic testing with the purported objective of empowering consumers to learn more about and manage their health. While understanding more about one's genetic make-up has intuitive appeal, concerns may arise if recipients of this information take inappropriate courses of action on the basis of not fully understanding (a) their test results, and/or (b) the complexity of genetics associated with certain phenotypes (Burton, 2015). In the following section we illustrate how failure to attend to the substantive aspects of validity, with an example that relates to questionnaire design, can preclude researchers and

practitioners from fully understanding how a particular psychological construct is related to salient health outcomes, and indeed (potentially) result in misdirection of intervention efforts.

Self-Efficacy in Health Behaviour Settings: A Case Study That Underscores the Importance of the Substantive Aspects of Validity

Within the field of health psychology (as well as other fields of psychology including education, sport, business, counselling psychology), the application of self-efficacy theory (Bandura, 1977, 1997) to understanding, and intervening, in relation to, *behavioural change* has been extensive. Embedded within a social-cognitive framework, self-efficacy is defined as a belief "in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura, 1997, p. 3), and is positioned as a major psychological determinant of a person's engagement in health-enhancing behaviours, along with the capacity to deal with adversity and persist in the face of considerable obstacles. Indeed, Bandura (1997) provided compelling evidence that a strong sense of self-efficacy can activate a range of biological processes that can both bolster human health and buffer against disease.

From a measurement perspective, Bandura (1997, 2006) repeatedly emphasized that self-efficacy beliefs are concerned with a person's confidence that they 'can do' a given behaviour and not whether they 'will do' a given behaviour. This distinction is important as the former corresponds to a belief about *capability*, whereas the latter represents a belief about *intention*. While this operationalization (with items framed by 'can do' questions) would certainly appear to address Messick's (1995) notion of content validity, in the form of both content relevance and representativeness, recent evidence points to potential concerns with the substantive aspects of validity that might exist within traditionally constructed self-efficacy instruments, especially those concerned with the self-regulation of health behaviours.

In a recent conceptual analysis of self-efficacy research within the field of health psychology, Williams and Rhodes (2016) explained that when people respond to traditional self-efficacy items/questionnaires, especially those concerned with the self-regulation of complex health behaviours (e.g., one's confidence to self-regulate regular physical activity behaviours in the face of various life challenges, one's confidence to maintain a healthy diet), their responses might inadvertently reflect motivation and not perceived capability as would be intended by the tenets of the underlying theory (Bandura, 1977, 1997). Specifically, in their critique, Williams and Rhodes (2016) drew from diverse sources of evidence, which suggest that measures derived from typical self-regulatory efficacy instruments may conflate capability with intention.