

Bruno D. Zumbo  
Anita M. Hubley *Editors*

# Understanding and Investigating Response Processes in Validation Research

# **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

<b>1</b>	<b>Response Processes in the Context of Validity: Setting the Stage.....</b>	<b>1</b>
	Anita M. Hubley and Bruno D. Zumbo	
<b>2</b>	<b>Response Processes and Measurement Validity in Health Psychology.....</b>	<b>13</b>
	Mark R. Beauchamp and Desmond McEwan	
<b>3</b>	<b>Contributions of Response Processes Analysis to the Validation of an Assessment of Higher Education Students' Competence in Business and Economics .....</b>	<b>31</b>
	Sebastian Brückner and James W. Pellegrino	
<b>4</b>	<b>Ecological Framework of Item Responding as Validity Evidence: An Application of Multilevel DIF Modeling Using PISA Data.....</b>	<b>53</b>
	Michelle Y. Chen and Bruno D. Zumbo	
<b>5</b>	<b>Putting Flesh on the Psychometric Bone: Making Sense of IRT Parameters in Non-cognitive Measures by Investigating the Social-Cognitive Aspects of the Items .....</b>	<b>69</b>
	Anita M. Hubley, Amery D. Wu, Yan Liu, and Bruno D. Zumbo	
<b>6</b>	<b>Some Observations on Response Processes Research and Its Future Theoretical and Methodological Directions .....</b>	<b>93</b>
	Mihaela Launeanu and Anita M. Hubley	
<b>7</b>	<b>A Model Building Approach to Examining Response Processes as a Source of Validity Evidence for Self-Report Items and Measures .....</b>	<b>115</b>
	Mihaela Launeanu and Anita M. Hubley	
<b>8</b>	<b>Response Processes and Validity Evidence: Controlling for Emotions in Think Aloud Interviews .....</b>	<b>137</b>
	Jacqueline P. Leighton, Wei Tang, and Qi Guo	

<b>9</b>	<b>Response Time Data as Validity Evidence: Has It Lived Up To Its Promise and, If Not, What Would It Take to Do So</b> .....	159
	Zhi Li, Jayanti Banerjee, and Bruno D. Zumbo	
<b>10</b>	<b>Observing Testing Situations: Validation as Jazz</b> .....	179
	Bryan Maddox and Bruno D. Zumbo	
<b>11</b>	<b>A Rationale for and Demonstration of the Use of DIF and Mixed Methods</b> .....	193
	José-Luis Padilla and Isabel Benítez	
<b>12</b>	<b>Cognitive Interviewing and Think Aloud Methods</b> .....	211
	José-Luis Padilla and Jacqueline P. Leighton	
<b>13</b>	<b>Some Thoughts on Gathering Response Processes Validity Evidence in the Context of Online Measurement and the Digital Revolution</b> .....	229
	Lara B. Russell and Anita M. Hubley	
<b>14</b>	<b>Longitudinal Change in Response Processes: A Response Shift Perspective</b> .....	251
	Richard Sawatzky, Tolulope T. Sajobi, Ronak Brahmhatt, Eric K.H. Chan, Lisa M. Lix, and Bruno D. Zumbo	
<b>15</b>	<b>Validating a Distractor-Driven Geometry Test Using a Generalized Diagnostic Classification Model</b> .....	277
	Benjamin R. Shear and Louis A. Roussos	
<b>16</b>	<b>Understanding Test-Taking Strategies for a Reading Comprehension Test via Latent Variable Regression with Pratt’s Importance Measures</b> .....	305
	Amery D. Wu and Bruno D. Zumbo	
<b>17</b>	<b>An Investigation of Writing Processes Employed in Scenario-Based Assessment</b> .....	321
	Mo Zhang, Danjie Zou, Amery D. Wu, Paul Deane, and Chen Li	
<b>18</b>	<b>National and International Educational Achievement Testing: A Case of Multi-level Validation Framed by the Ecological Model of Item Responding</b> .....	341
	Bruno D. Zumbo, Yan Liu, Amery D. Wu, Barry Forer, and Benjamin R. Shear	
<b>19</b>	<b>On Models and Modeling in Measurement and Validation Studies</b> .....	363
	Bruno D. Zumbo	

# 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 Brahmhatt** 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*<sup>1</sup> (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

---

<sup>1</sup>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](mailto:anita.hubley@ubc.ca); [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca)

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.<sup>2</sup> Thus, if different respondents respond using different processes, then content validity may differ among those respondents despite the items being the same.

---

<sup>2</sup>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 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](https://doi.org/10.1111/j.1525-1497.2005.0258.x).
- Brückner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multi-level 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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.
- Ivrie, 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](https://doi.org/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. *Applied Psychological Measurement*, *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](mailto:mark.beauchamp@ubc.ca); [desi.mcewan@ubc.ca](mailto:desi.mcewan@ubc.ca)

© Springer International Publishing AG 2017

B.D. Zumbo, A.M. Hubble (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 administer 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.

From a theoretical perspective, Bandura (1978, 1997, 2004) has repeatedly emphasized over the years that, from a temporal perspective, self-efficacy beliefs causally precede outcome expectations but are not influenced by outcome expectations. Balanced against this theoretical postulate, Williams and Rhodes (2016) summarized the results of a series of experimental studies whereby health-related outcome expectations were, contrary to the theoretical tenets of self-efficacy theory, found to causally influence self-efficacy beliefs. As a complement to this work, Williams and Rhodes also drew from the results of thought-listing research that involved asking participants (using an open-ended response format) to consider their answers to a series of self-efficacy items and explain “the main reasons why you are generally confident or unconfident you can overcome these barriers and engage in regular physical activity” (Rhodes & Blanchard, 2007, p. 763). Through this research, various motivational factors (e.g., expectations of improved health, enjoyment) were identified by participants as explanations for their responses to self-efficacy items. Finally, Williams and Rhodes drew from a series of psychometric studies, in which traditional self-efficacy items were augmented by efforts to hold motivation constant and compared to the original self-efficacy items. Specifically, in studies by Rhodes and colleagues (Rhodes & Blanchard, 2007; Rhodes & Courneya, 2003, 2004), the relative predictive utility of responses to traditional self-efficacy items were compared with responses to items that also included the qualifier ‘if you really wanted to’ at the end of those self-efficacy items (thus, holding motivation constant). For example, “how confident are you that you can exercise when tired [if you really wanted to]”? What they found was the mean scores of participants’ responses went up in the augmented measures, when compared to responses to traditional self-efficacy items, and also the correlations with behavioural intention measures became notably weaker. When taken together, what this body of research suggests is that traditional measures of self-efficacy (unintentionally) tap into an assessment of motivation, at least to some extent (see Williams & Rhodes, 2016, for a full discussion of this debate).

As an explanation for how and or why this happens, it is worth reflecting on how people interpret language in a colloquial versus a literal sense. In a literal sense, questions that ostensibly appear to reflect ‘a person’s confidence that s/he *can do* behaviour X’ would seem to align well with Bandura’s definition of capability, and indeed whether the person perceives that s/he can perform the given behaviour. However, drawing from reasoning by Kirsch (1995), Williams and Rhodes (2016) contended that when people are asked whether they can perform health behaviours (within a given time frame), such as eating a healthy diet or abstaining from drinking alcohol for a month, they may often query whether the question/item is asking them whether they are actually physically capable of performing the given behaviour or whether they likely *will* eat a healthy diet or refrain from drinking within that time frame. Consider this simple thought experiment: if a friend asks ‘can you go to the cinema tomorrow?’ would you interpret this as reflecting whether you are *physically capable of* going to the movies or whether you *want to* go to the movies?

Unfortunately, despite the persuasiveness of Williams and Rhodes’ (2016) argument that traditional measures of self-efficacy might inadvertently measure



motivation/intention at least to some extent, research has yet to examine what people are thinking about ('in situ') while they are responding to traditional self-efficacy items/questionnaires. As explained by Beauchamp (2016), one methodological approach that has the potential to shed light on this question corresponds to the use of 'think-aloud' protocols (Oremus et al., 2005; Willis, 2005). Think aloud protocols represent a form of cognitive interviewing whereby research participants explain exactly what they are thinking about while they are completing questionnaire items. Should data derived from such an approach provide support for Williams and Rhodes' critical contention that self-efficacy questionnaires (and the items subsumed within them) do in fact assess intention (*will-do motivation* rather than *can-do capability*), this would have non-trivial implications for understanding the predictive utility of the self-efficacy construct as well as health promotion interventions that have developed various initiatives on the basis of findings that link self-regulatory efficacy beliefs with various health-enhancing behaviours.

Specifically, should responses to self-efficacy items unintentionally tap into measures of motivation, then any conclusions derived from correlations from such measures in relation to health behaviour outcomes would mask any insights in terms of whether 'perceived capability' is driving the effect (i.e., explaining the majority of variance) or whether 'intention' is the most salient predictor. This issue ties directly to Messick's (1995) articulation concerning the external aspects of validity, notably the extent to which test scores from a focal construct relate to a theorized target outcome (in this case health behaviour change). Indeed, it is conceivable that self-efficacy may have been given undue credit in its capacity to predict various health behaviours (at least in the context of self-regulatory health behaviours). If this is the case, this would also have noteworthy implications from a consequential validity perspective as well. For instance, on the basis of consistent findings linking self-efficacy beliefs to better engagement in health behaviours, such as improved physical activity (McAuley & Blissmer, 2000), healthy eating (e.g., Anderson, Winnett, & Wojcik, 2007), smoking cessation (Gwaltney, Metrik, Kahler, & Shiffman, 2009), and safe sex (e.g., Sayles et al., 2006), an extensive range of health behaviour interventions have been developed over the past few years with a primary goal of bolstering participants' self-regulatory efficacy beliefs (e.g., Bryan, Robbins, Ruiz, & O'Neill, 2006; Elfeddali, Bolman, Candel, Wiers, & De Vries, 2012; Luszczynska, & Tryburcy, 2008; Luszczynska, Tryburcy, & Schwarzer, 2007). In short, if responses to self-efficacy items conflate or merge perceptions of ability with conceptions of motivation then, from an applied/interventionist perspective, one might be precluded from understanding and thereafter targeting the most relevant psychological state/cognition.

In sum, in order to guide effective interventions it is critical that one is able to disentangle measures of capability from motivation, for any intervention to be effective (Beauchamp, 2016). When taken together, and as this example illustrates, it is absolutely critical that attention is directed to the substantive aspects of validity and ensuring that participants' response processes in relation to measurement devices (i.e., questionnaires) align with what is purported to be measured. Unfortunately, this critical methodological 'step' is often overlooked, with researchers simply



cobbling together a collection of items that they believe reflect the content area and then subjecting participants' responses to psychometric analyses (e.g., factor analysis); the results of this (mal)practice can be a failure to sufficiently test the underpinning theoretical framework and, worse, have a negative impact on individuals' health (through inappropriately targeted interventions).

## **Methods to Support the Validity of Response Processes in Health Psychology**

With this in mind, in this section we provide an overview of different methodological approaches that are specifically concerned with optimizing the substantive aspects of validity in relation to participants' response processes within the field of health psychology. Specifically, we discuss the use of think aloud protocols (e.g., Gadermann, Guhn, M., & Zumbo, 2011) implicit measures (e.g., Greenwald, Poehlman, Uhlmann & Banaji, 2009), as well as behavioural measures (e.g., Bassett-Gunter, Latimer-Cheung, Martin Ginis, & Castelhana, 2014) in the context of health psychology research.

### ***Think Aloud Protocols***

Think aloud protocols can be conducted individually or in group-based formats. In each case, participants are instructed to complete a copy of the initial questionnaire independently and following this, a series of questions are used in order to prompt participants to discuss questionnaire items in terms of the instructions, response format, and wording of items. Think aloud protocols have also been referred to as retrospective verbalization (Ericsson & Simon, 1980). Typical questions might include those such as (a) "What, in your own words, does the question mean to you?", (b) "Did the answer choices include your answer?", (c) "Did you understand how to answer the questions?", and (d) "Did the questionnaire leave anything out you felt was important?" (Oremus et al., 2005; Willis, 2005). Conversations with participants are recorded and subject to content analysis. In our own previous work using this methodology (e.g., Morton et al., 2011; Sylvester et al., 2014), we have often used a qualitative constant comparison approach (Strauss & Corbin, 1998) to identify and code components of items with which participants raise concerns. The subsequent analyses focus on identifying problematic and/or alternative interpretations of items, and then reworking/rephrasing them accordingly. This process is followed on an iterative basis with additional participants/groups until no further suggestions for revision emerge. In working with specific populations, such as young children, additional *a priori* steps can be utilized such as ensuring the readability of items is targeted at the appropriate developmental stage of the target

population (Morton et al., 2011). This can be done by modifying and/or simplifying item structure, as well as the accompanying instructions, to ensure that items reflect an appropriate reading ease score (Flesch, 1948).

In the field of health psychology, the use of think aloud protocols has increasingly been used to good effect. For example, in one study that examined what people are thinking about when they answer theory of planned behavior (Ajzen, 1991) questionnaires, French et al. (2007) found that some participants had problems with information retrieval associated with some of the items, as well as answering different questions than those that were intended by the researchers. In addition to identifying potential problems with item construction/wording, think aloud approaches can also be extremely useful in providing validity evidence related to response processes when questionnaires have been adapted or revised for a new population. In a series of studies by Gadermann and colleagues (Gadermann, Schonert-Reichl, & Zumbo, 2010, Gadermann et al. 2011), the authors sought to adapt a well-known measure of life satisfaction, namely the Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985), for use with children. In so doing, they developed the Satisfaction with Life Scale Adapted for Children (SWLS-C; Gadermann et al., 2010), and initially provided evidence for measurement reliability, unidimensionality (as per the original SWLS), as well as invariance across gender, first language learned at home, and grade level. In addition, and of direct relevance to the current chapter, Gadermann et al. (2011) also conducted a think aloud study related to the items subsumed within the SWL-C. While the study revealed that most children had no difficulty with item interpretation, the use of a think aloud protocol also highlighted two distinct strategies that children used to base their responses. Specifically, children displayed the use of both *absolute* as well as *relative* strategies in responding to items (Gadermann et al., 2011). Absolute strategies reflected children explaining the overall presence or absence of an event tied directly to their appraisal of life satisfaction (“I am happy with my life because I have a *really* caring family” [emphasis added]). In contrast, relative strategies were reflected through comparative appraisals within children’s responses (“I want to have *more* friends” [emphasis added]).

In a separate study by Morton et al. (2011) that was designed to assess parenting behaviours in the context of adolescent health promotion, a think aloud protocol was utilized to ensure that items were relevant and interpretable by both adolescents and their parents. As a result of the iterative think aloud protocol, which was delivered via focus group format, some items were modified slightly in terms of wording. In addition, a few adolescents perceived some of the items to be difficult to comprehend and these items were omitted. Finally, changes were made to the verbal anchors affixed to each response option. Initially, the response format was a 0-4 scale which asked about the *frequency* of parenting behaviors. However, some respondents discussed that the “frequency” response was difficult to comprehend for some items. As one participant noted, “It would be better to have ‘agree’ or ‘disagree’ because ‘frequently’ is a timely basis and not all of these are done every day; they don’t *always* do it but it’s still there” (Morton et al., 2011, p. 704). As a result, the final version of the authors’ parenting instrument comprised a *strongly disagree* to

*strongly agree* response format (Morton et al., 2011). In sum, by involving participants in a thorough examination of their response processes, issues related to scaling (i.e., use of anchors) and item wording (clarity, phraseology, structure) can be enhanced to ensure that they directly align with the theorized substance that underpins the given measure. When taken together, the use of think-aloud protocols represents a simple, cost-effective, but highly efficacious means of examining research participants' response processes to ensure that they reflect the target cognition of interest.

### ***Implicit Measures***

Although the use of questionnaires represents the most pervasive means of assessing psychological processes within the field of health psychology, other methodological approaches exist that have the potential to overcome some of the limitations of self-report measures, especially those that deal with sensitive issues (Greenwald et al., 2009). Drawing from a dual-systems model perspective (Strack & Deutsch, 2004), researchers in health psychology have increasingly made use of "implicit" or "indirect" measures of psychological processes alongside (or instead of) traditional explicit questionnaire measures. Implicit measures are designed to provide an assessment of a psychological state or cognition in which the "outcome [measure] functions as an index of an attitude or cognition despite the fact that participants are unaware of the impact of the attitude or cognition on the outcome, are not aware of the attitude or outcome, or have no control over the outcome" (de Houwer, 2006, p. 12). The most widely used implicit measure in psychology is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), although many other such measures have been developed and used within psychology in recent years.

In terms of the external aspect of validity (cf. Messick, 1995) responses to implicit measures have been found to be better predictors of spontaneous behaviours, whereas explicit measures tend to outperform implicit measures in more deliberate behaviour (Gawronski & De Houwer, 2014). From a response process perspective, one of the advantages of implicit measures is that they do not require people to control their responses to items/stimuli and do not require introspection that is typically required for explicit assessments (Gawronski & De Houwer, 2014).

Proponents of implicit measures have also argued that such measures are less susceptible to respondent biases and social desirability motives than traditional explicit measurement approaches (Gawronski & Bodenhausen, 2006). In recognition of the contribution that both automatic and reflective psychological processes likely play in the prediction of health behaviours, Keatley et al. (2012) sought to examine the relations between implicit and explicit measures of autonomous and controlled motivation and 20 health-related behaviours. The results revealed that, in general, the prediction of health behaviours was more effective through explicit (i.e., reflective) measures of motivational constructs, although the results also provided some support for the predictive utility of implicit measures.

Balanced against these potential strengths, others have raised concerns about the psychometric properties of implicit measures (Blanton, Jaccard, & Burrows, 2015a; Blanton, Jaccard, Strauts, Mitchell, & Tetlock, 2015b). As Blanton et al. (2015b) note, implicit measures are scored on an arbitrary metric that “have yet to be systematically mapped onto true scores on the underlying dimension .... As such, researchers cannot definitively link the degree of behavioral bias to any specific IAT score, and it follows that they also cannot use the distribution of IAT scores to infer either the prevalence or the average magnitude of behavioral bias in any given group” (p. 1). Nevertheless, recent psychometric work within this area has concertedly sought to examine the distributional properties of implicit judgements with a view to making implicit metrics more meaningful (Blanton et al. 2015b). Specifically, Blanton et al. conducted a secondary analysis of previous studies that examined implicit prejudice among Americans, and observed that implicit measures of prejudice tend to be “right biased” and overestimate the prevalence of biases in a given population. When taken together, the use of implicit measures offers considerable advantages over traditional (explicit) measures. As this field of assessment develops, it will become critical to ascertain that the implicit/automatic responses of participants, as Blanton et al. (2015b) note, accurately map on to the underlying psychological construct.

### *Behavioural Measures*

In addition to extensively utilized explicit measures (i.e., questionnaires), and the more recent contributions of implicit measures, other methodological approaches have been utilized that are specifically concerned with ensuring that assessments of various psychological processes reflect what is intended to be measured (i.e., high substantive validity; Messick, 1995). In this section, we discuss two behavioural measures that represent a viable alternative to questionnaire-based assessments (in some instances). As with explicit and implicit measures, both involve examining participants’ response processes but, as advocates of each approach contend, are posited to tap into psychological processes more directly.

The first corresponds to the use of eye-tracking assessments to assess cognitive processes in relation to health-information stimuli. The use of advertising and health messaging has been a pervasive method of seeking to promote population-level behaviour change (Noar, Benac, & Harris, 2007). Indeed, advertising has been used to encourage most health-enhancing behaviours (e.g., physical activity, diet) as well as discourage health-compromising behaviours (e.g., smoking cessation, alcohol abstinence, safe-sex, drug avoidance, stop texting and driving). Research in this area has utilized eye-tracking technology to examine viewer attention and cognitive processing of gain-framed (i.e., emphasizing the benefits of engaging in a health behaviour) versus loss-framed (i.e., emphasizing the risks of not engaging in a health behaviour) advertisements, via examination of participant fixations, dwell time, and recall of those messages (O’Malley & Latimer-Cheung 2013). While research on

the uptake and engagement of health-behaviours in relation to advertising/messaging via self-report questionnaires has been extensive (e.g., Anderson, De Bruijn, Angus, Gordon, & Hastings, 2009), such approaches are often subject to memory decay and social desirability biases (Prince et al., 2008). What eye-tracking technology offers is the opportunity to (objectively) examine where people are directing their attention when confronted with health-related messages. That is, as with the use of implicit measures, the use of eye tracking technology does not require introspection among respondents with regard to their interpretation of items/questions, and instead allows researchers to examine where participant attention is directed in situ (e.g., at the same time they are reading and evaluating any health-promotion materials); thus deriving potentially stronger measures of the focal construct.

Several studies have shown that gain-framed health messages elicit greater attention (measured by dwell time on a message; e.g., Bassett-Gunter et al., 2014; O'Malley & Latimer-Cheung 2013) and cognitive processing (measured by message recall and message-relevant thoughts; O'Keefe & Jensen, 2008) compared to loss-framed messages. Although, as highlighted by O'Malley and Latimer-Cheung (2013), research has yet to establish the role of cognitive processing (via eye-tracking technology) as a mediator of the relationships between variously framed messages and health behaviour outcomes, this approach appears to show promise in objectively measuring health-related behaviour. Indeed, such an approach is not hampered by potential problems with memory decay, social desirability biases, or the requirement for participant introspection, but allows researchers to quantify in real time how much attention is directed to the target stimuli.

The second behavioural approach that represents a viable alternative to the use of questionnaires, in assessing psychological processes, corresponds to the use of 'performance-based' (or ability-based) assessments. An excellent example of this corresponds to the assessment of emotional intelligence (EI). EI refers to a person's ability to perceive, use, understand, and manage emotions in order to facilitate social functioning (Mayer, Roberts, & Barsade, 2008; Mayer, Salovey, & Caruso, 2004, 2008). This definition reflects the multi-dimensionality of EI and, in particular, the four core emotion-related 'skills' that are purported to constitute EI (Mayer, et al., 2004; Mayer, Salovey, & Caruso, 2008). These skills include the *perception of emotion*, which reflects a person's ability to recognize various emotions in others as well as oneself, through their body language, interactions, oral communication, facial expressions and so forth. The *use of emotions* reflects a person's ability to generate different emotions in oneself and others (e.g., happiness, pride) in order to achieve some desired outcome (e.g., improved contributions to a group). *Understanding emotions* reflects a meta-cognitive process (i.e., cognitions about cognitions) that reflects a person's knowledge about the emotions that they, and others, experience. This might involve an understanding of how various emotions emerge in the first place (i.e., antecedents), as well as the likely downstream effects that these different emotions (i.e., consequences) might have on oneself and others. Finally, the *management of emotions* reflects one's ability to control one's own emotions, as well as the ability to regulate the emotions of other people.

The importance of EI has been demonstrated across a vast array of life contexts, including education (Williams, 2008), health care (Arora, et al., 2010), business (Ashkanasy & Humphrey, 2011), and sport (Crombie, Lombard, & Noakes, 2009). In the context of health, emotional intelligence appears to be implicated in supporting improvements in physical, psychosomatic, and mental health (Schutte, Malouff, Thorsteinsson, Bhullar, & Rooke, 2007). Indeed, emotional intelligence appears to be important across a range of health-related variables at the individual level (e.g., health-related coping behaviours; Saklofske, Austin, Galloway, & Davidson, 2007), as well as in dyadic relationships (e.g., caring behaviour among nurses; Rego, Godinho, McQueen, & Cunha, 2010) and group interactions (e.g., medical teams' levels of cohesion; Quidbach & Hansenne, 2009).

Of direct relevance to the current chapter, there has been considerable debate in terms of how EI should be assessed, as well as an appraisal of the reliability and validity evidence in support of these different approaches. There have been several efforts to assess EI through various self-report inventories such as Bar-On's (1997) *Emotional Quotient Inventory* and Schutte and colleagues' (1998) self-report EI test (*SEIT*). However, as highlighted by Brackett and Mayer (2003), these self-report measures appear to be subject to non-trivial self-report bias, as well as a lack of discriminant validity in relation to well-established personality measures. Indeed, if due to the very nature of a psychological construct, any self-report assessment of that variable (e.g., questionnaire assessments) results in responses that do not align with what is purported to be assessed, this has non-trivial implications for understanding the nature of that construct. If EI is theorized to be conceptually distinct from personality, then measures of EI and personality should be unrelated. In contrast to these self-report measures, Mayer, Salovey, and Caruso (2002) developed the Mayer Salovey Caruso Emotional Intelligence Test (MSCEIT), which represents an objective assessment of a participant's capacity to solve emotion-laden problems. As highlighted by Duncan, Latimer-Cheung, and Brackett, (2014), the MSCEIT is considered an objective test of EI, as scores are determined in relation to normative and expert samples and, in particular, "how correct the answers are vis-à-vis the norms generated from [these] normative or expert samples" (p. 6). The MSCEIT quantifies participants' abilities to perceive, use, understand, and manage emotions to facilitate social functioning. As an example of the behavioural nature of the MSCEIT assessment procedures, perceptions of emotion are assessed by having people rate how much of an emotion is exhibited in people's faces as well as in pictures and landscapes (Brackett & Mayer, 2003). In contrast, the management of emotion is assessed by having people choose effective ways to manage different emotions, as well as the emotions of other people, in a range of private and interpersonal situations (Brackett & Mayer, 2003; Mayer et al., 2003). Scores derived from the MSCEIT have been found to display good reliability and factorial validity (Mayer et al., 2003) along with discriminant validity in relation to measures of personality (Brackett & Mayer, 2003) as well as measures of intelligence (Brackett, Mayer, & Warner, 2004). It is beyond the scope of this chapter to provide a comprehensive overview of the EI construct. However, what this body of research illustrates is that, in instances whereby some psychological 'ability' represents the



focus of enquiry (e.g., resilience), the use of objective ability-based assessments might represent a viable alternative to assessing that construct through typical self-report questionnaire-based approaches.

## Conclusion

As several scholars (e.g., Downing, 2003; Marsh, 1997; Messick, 1995) have noted over the years, validation is an on-going process. We take the view of Messick (1995) that construct validity is comprised of multiple aspects and, in line with Downing (2003), that “validity is never assumed and is an ongoing process of hypothesis generation, data collection and testing, critical evaluation and logical inference” (p. 831). In this chapter, we focused on the importance of considering the substantive aspects of validity and, in particular, how reliable and valid assessments of response processes can provide the foundation for the remaining aspects of validity, especially the external and consequential components described by Messick (1995). That is, if the responses of participants to a given psychological assessment do not align with what is posited to be measured, then this has non-trivial implications that include a failure to fully understand the nature of the psychological construct, as well as interventions that are sub-optimal and even problematic. We also sought to provide an overview of some methodological approaches that can be used to enhance the substantive aspects of measurement validity and, in particular, the validity of response processes in health psychology research.

## References

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, 179–211. doi:10.1016/0749-5978(91)90020-T.
- Anderson, E. S., Winett, R. A., & Wojcik, J. R. (2007). Self-regulation, self-efficacy, outcome expectations, and social support: Social cognitive theory and nutrition behavior. *Annals of Behavioral Medicine*, 34, 304–312. doi:10.1007/BF02874555.
- Anderson, P., De Bruijn, A., Angus, K., Gordon, R., & Hastings, G. (2009). Impact of alcohol advertising and media exposure on adolescent alcohol use: A systematic review of longitudinal studies. *Alcohol and Alcoholism*, 44, 229–243. doi:10.1093/alcalc/agn115.
- Arora, S., Ashrafian, H., Davis, R., Athanasiou, T., Darzi, A., & Sevdalis, N. (2010). Emotional intelligence in medicine: A systematic review through the context of the ACGME competencies. *Medical Education*, 44, 749–764. doi:10.1111/j.1365-2923.2010.03709.x.
- Ashkanasy, N. M., & Humphrey, R. H. (2011). Current emotion research in organizational behavior. *Emotion Review*, 3, 214–224. doi:10.1177/1754073910391684.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191–215. doi:10.1037/0033-295X.84.2.191.
- Bandura, A. (1978). Reflections on self-efficacy. *Advances in Behaviour Research and Therapy*, 1, 237–269. doi:10.1016/0146-6402(78)90012-7.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W.H. Freeman and Co.

- Bandura, A. (2004). Health promotion by social cognitive means. *Health Education and Behavior*, 3, 143–164. doi:[10.1177/1090198104263660](https://doi.org/10.1177/1090198104263660).
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. C. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Greenwich, UK: Information Age.
- Bar-On, R. (1997). *Bar-On emotional quotient inventory: Technical manual*. Toronto, ON: Multi-Health Systems.
- Bassett-Gunter, R. L., Latimer-Cheung, A. E., Martin Ginis, K. A., & Castelhana, M. (2014). I spy with my little eye: Cognitive processing of framed physical activity messages. *Journal of Health Communication*, 19, 676–691. doi:[10.1080/10810730.2013.837553](https://doi.org/10.1080/10810730.2013.837553).
- Beauchamp, M. R. (2016). Disentangling motivation from self-efficacy: Implications for measurement, theory-development, and intervention. *Health Psychology Review*, 10, 129–132. doi:[10.1080/17437199.2016.1162666](https://doi.org/10.1080/17437199.2016.1162666).
- Beauchamp, M. R., Bray, S. R., Eys, M. A., & Carron, A. V. (2002). Role ambiguity, role efficacy, and role performance: Multidimensional and mediational relationships within interdependent sport teams. *Group Dynamics: Theory, Research, and Practice*, 6, 229–242. doi:[10.1037//1089-2699.6.3.229](https://doi.org/10.1037//1089-2699.6.3.229).
- Beauchamp, M. R., Barling, J., Li, Z., Morton, K. L., Keith, S. E., & Zumbo, B. D. (2010). Development and psychometric properties of the Transformational Teaching Questionnaire. *Journal of Health Psychology*, 15, 1123–1134. doi:[10.1177/1359105310364175](https://doi.org/10.1177/1359105310364175).
- Blanton, H., Jaccard, J., & Burrows, C. N. (2015a). Implications of the implicit association Test D-transformation for psychological assessment. *Assessment*, 22, 429–440. doi:[10.1177/1073191114551382](https://doi.org/10.1177/1073191114551382).
- Blanton, H., Jaccard, J., Strauts, E., Mitchell, G., & Tetlock, P. E. (2015b). Toward a meaningful metric of implicit prejudice. *Journal of Applied Psychology*, 100, 1468–1481. doi:[10.1037/a0038379](https://doi.org/10.1037/a0038379).
- Brackett, M. A., & Mayer, J. D. (2003). Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. *Personality and Social Psychology Bulletin*, 29, 1147–1158. doi:[10.1177/0146167203254596](https://doi.org/10.1177/0146167203254596).
- Brackett, M. A., Mayer, J. D., & Warner, R. M. (2004). Emotional intelligence and its relation to everyday behaviour. *Personality & Individual Differences*, 36, 1387–1402. doi:[10.1016/S0191-8869\(03\)00236-8](https://doi.org/10.1016/S0191-8869(03)00236-8).
- Bryan, A., Robbins, R. N., Ruiz, M. S., & O’Neill, D. (2006). Effectiveness of an HIV prevention intervention in prison among African Americans, Hispanics, and Caucasians. *Health Education and Behavior*, 33, 154–177. doi:[10.1177/1090198105277336](https://doi.org/10.1177/1090198105277336).
- Burton, A. (2015). Are we ready for direct-to-consumer genetic testing? *The Lancet Neurology*, 14, 138–139. doi:[10.1016/S1474-4422\(15\)70003-7](https://doi.org/10.1016/S1474-4422(15)70003-7).
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7, 309–319.
- Crombie, D., Lombard, C., & Noakes, T. (2009). Emotional intelligence scores predict team sports performance in a national cricket competition. *International Journal of Sports Science and Coaching*, 4, 209–224. doi:[10.1260/174795409788549544](https://doi.org/10.1260/174795409788549544).
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- De Houwer, J. (2006). What are implicit measures and why are we using them? In W. W. Reinout & A. W. Stacy (Eds.), *Handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks, CA: Sage Publications. doi:[10.4135/9781412976237.n2](https://doi.org/10.4135/9781412976237.n2).
- De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, 135, 347–368. doi:[10.1037/a0014211](https://doi.org/10.1037/a0014211).
- Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with Life Scale. *Journal of Personality Assessment*, 49, 71–75.
- 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](https://doi.org/10.1046/j.1365-2923.2003.01594.x).



- Duncan, L., Latimer-Cheung, A., & Brackett, M. A. (2014). Emotional intelligence: A framework for examining emotions in sport and exercise groups. In M. R. Beauchamp & M. A. Eys (Eds.), *Group dynamics in exercise and sport psychology* (2nd ed., pp. 3–20). New York, NY: Routledge.
- Elfeddali, I., Bolman, C., Candel, M. J., Wiers, R. W., & De Vries, H. (2012). The role of self-efficacy, recovery self-efficacy, and preparatory planning in predicting short-term smoking relapse. *British Journal of Health Psychology*, *17*, 185–201. doi:10.1111/j.2044-8287.2011.02032.x.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233. doi:10.1037/h0057532.
- French, D. P., Cooke, R., McLean, N., & Williams, M. (2007). What do people think about when they answer theory of planned behaviour questionnaires?: A ‘think aloud’ study. *Journal of Health Psychology*, *12*, 672–687. doi:10.1177/1359105307078174.
- Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction with Life Scale adapted for children: A focus on cognitive process. *Social Indicators Research*, *100*, 37–60. doi:10.1007/s11205-010-9603-x.
- Gadermann, A. M., Schonert-Reichl, K. A., & Zumbo, B. D. (2010). Investigating validity evidence of the satisfaction with Life Scale adapted for children. *Social Indicators Research*, *96*, 229–247. doi:10.1007/s11205-009-9474-1.
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731. doi:10.1037/0033-2909.132.5.692.
- Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 283–310). New York, NY: Cambridge University Press.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. doi:10.1037/0022-3514.74.6.1464.
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41. doi:10.1037/a0015575.
- Gwaltney, C. J., Metrik, J., Kahler, C. W., & Shiffman, S. (2009). Self-efficacy and smoking cessation: A meta-analysis. *Psychology of Addictive Behaviors*, *23*, 56–66. doi:10.1037/a0013529.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. doi:10.1080/10705519909540118.
- Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology*, *123*, 207–215. doi:10.1080/00221309.1996.9921273.
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*, 219–230. doi:10.1007/s11205-011-9843-4.
- Johnston, M., French, D. P., Bonetti, D., & Johnston, D. W. (2004). Assessment and measurement in health psychology. In S. Sutton, A. Baum, & M. Johnston (Eds.), *The Sage handbook of health psychology* (pp. 288–323). London, UK: Sage.
- Keatley, D., Clarke, D. D., & Hagger, M. S. (2012). The predictive validity of implicit measures of self-determined motivation across health-related behaviours. *British Journal of Health Psychology*, *18*, 2–17. doi:10.1111/j.2044-8287.2011.02063.x.
- Kirsch, I. (1995). Self-efficacy and outcome expectancy: A concluding commentary. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment: Theory, research, and application* (pp. 341–345). New York, NY: Plenum Press.
- Luszczynska, A., & Tryburcy, M. (2008). Effects of a self-efficacy intervention on exercise: The moderating role of diabetes and cardiovascular diseases. *Applied Psychology*, *57*, 644–659. doi:10.1111/j.1464-0597.2008.00340.x.

- Luszczynska, A., Tryburcy, M., & Schwarzer, R. (2007). Improving fruit and vegetable consumption: A self-efficacy intervention compared with a combined self-efficacy and planning intervention. *Health Education Research*, 22, 630–638. doi:[10.1093/her/cyl133](https://doi.org/10.1093/her/cyl133).
- Marsh, H. W. (1997). The measurement of physical self-concept: A construct validation approach. In K. R. Fox (Ed.), *The physical self: From motivation to well-being* (pp. 27–58). Champaign, IL: Human Kinetics.
- Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology*, 59, 507–536. doi:[10.1146/annurev.psych.59.103006.093646](https://doi.org/10.1146/annurev.psych.59.103006.093646).
- Mayer, J. D., Salovey, P., & Caruso, D. R. (2002). *MSCEIT user's manual*. Toronto, ON: Multi-Health Systems.
- Mayer, J. D., Salovey, P., & Caruso, D. R. (2004). Emotional intelligence: Theory, findings, and implications. *Psychological Inquiry*, 15, 197–215. doi:[10.1207/s15327965pli1503\\_02](https://doi.org/10.1207/s15327965pli1503_02).
- Mayer, J. D., Salovey, P., & Caruso, D. R. (2008). Emotional intelligence: New ability or eclectic traits? *American Psychologist*, 63, 503–517. doi:[10.1037/0003-066X.63.6.503](https://doi.org/10.1037/0003-066X.63.6.503).
- Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion*, 3, 97–105. doi:[10.1037/1528-3542.3.1.97](https://doi.org/10.1037/1528-3542.3.1.97).
- McAuley, E., & Blissmer, B. (2000). Self-efficacy determinants and consequences of physical activity. *Exercise and Sport Sciences Reviews*, 28, 85–88.
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195–244. doi:[10.2466/pr0.1990.66.1.195](https://doi.org/10.2466/pr0.1990.66.1.195).
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 12–103). New York, NY: Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749. doi:[10.1037/0003-066X.50.9.741](https://doi.org/10.1037/0003-066X.50.9.741).
- Morton, K. L., Barling, J., Mâsse, L., Rhodes, R., Zumbo, B. D., & Beauchamp, M. R. (2011). The application of transformational leadership theory to parenting: Questionnaire development and implications for adolescent self-regulatory efficacy and life satisfaction. *Journal of Sport and Exercise Psychology*, 33, 688–709.
- Noar, S. M., Benac, C. N., & Harris, M. S. (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin*, 133, 673–693. doi:[10.1037/0033-2909.133.4.673](https://doi.org/10.1037/0033-2909.133.4.673).
- O'Keefe, D. J., & Jensen, J. D. (2008). Do loss-framed persuasive messages engender greater message processing than do gain-framed messages? A meta-analytic review. *Communication Studies*, 59, 51–67. doi:[10.1080/10510970701849388](https://doi.org/10.1080/10510970701849388).
- O'Malley, D. A., & Latimer-Cheung, A. E. (2013). Gaining perspective: The effects of message frame on viewer attention to and recall of osteoporosis prevention print advertisements. *Journal of Health Psychology*, 18, 1400–1410. doi:[10.1177/1359105312456323](https://doi.org/10.1177/1359105312456323).
- Oremus, M., Cosby, J. L., & Wolfson, C. (2005). A hybrid qualitative method for pretesting questionnaires: The example of a questionnaire to caregivers of Alzheimer disease patients. *Research in Nursing Health*, 28, 419–430. doi:[10.1002/nur.20095](https://doi.org/10.1002/nur.20095).
- Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Gorber, S. C., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 5, 56–70. doi:[10.1080/02701367.2000.11082780](https://doi.org/10.1080/02701367.2000.11082780).
- Quoidbach, J., & Hansenne, M. (2009). The impact of trait emotional intelligence on nursing team performance and cohesiveness. *Journal of Professional Nursing*, 25, 23–29. doi:[10.1016/j.profnurs.2007.12.002](https://doi.org/10.1016/j.profnurs.2007.12.002).
- Rego, A., Godinho, L., McQueen, A., & Cunha, M. P. (2010). Emotional intelligence and caring behaviour in nursing. *The Service Industries Journal*, 30, 1419–1437. doi:[10.1080/02642060802621486](https://doi.org/10.1080/02642060802621486).
- Rhodes, R. E., & Blanchard, C. M. (2007). What do confidence items measure in the physical activity domain? *Journal of Applied Social Psychology*, 37, 759–774. doi:[10.1111/j.1559-1816.2007.00184.x](https://doi.org/10.1111/j.1559-1816.2007.00184.x).

- Rhodes, R. E., & Courneya, K. S. (2003). Self-efficacy, controllability, and intention in the theory of planned behaviour: Measurement redundancy or causal independence? *Psychology and Health, 18*, 79–91. doi:[10.1080/0887044031000080665](https://doi.org/10.1080/0887044031000080665).
- Rhodes, R. E., & Courneya, K. S. (2004). Differentiating motivation and control in the theory of planned behaviour. *Psychology, Health, and Medicine, 9*, 205–215. doi:[10.1080/13548500410001670726](https://doi.org/10.1080/13548500410001670726).
- Saklofske, D. H., Austin, E. J., Galloway, J., & Davidson, K. (2007). Individual difference correlates of health-related behaviours: Preliminary evidence for links between emotional intelligence and coping. *Personality and Individual Differences, 42*, 491–502. doi:[10.1016/j.paid.2006.08.006](https://doi.org/10.1016/j.paid.2006.08.006).
- Sayles, J. N., Pettifor, A., Wong, M. D., MacPhail, C., Lee, S. J., Hendriksen, E., et al. (2006). Factors associated with self-efficacy for condom use and sexual negotiation among South African youth. *Journal of Acquired Immune Deficiency Syndromes, 43*, 226–233. doi:[10.1097/01.qai.0000230527.17459.5c](https://doi.org/10.1097/01.qai.0000230527.17459.5c).
- Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., et al. (1998). Development and validation of a measure of emotional intelligence. *Personality and Individual Differences, 25*, 167–177. doi:[10.1016/S0191-8869\(98\)00001-4](https://doi.org/10.1016/S0191-8869(98)00001-4).
- Schutte, N. S., Malouff, J. M., Thorsteinsson, E. B., Bhullar, N., & Rooke, S. E. (2007). A meta-analytic investigation of the relationship between emotional intelligence and health. *Personality and Individual Differences, 42*, 921–933. doi:[10.1016/j.paid.2006.09.003](https://doi.org/10.1016/j.paid.2006.09.003).
- Schwartz, B. (2015). *Why we work*. New York, NY: Simon and Schuster.
- Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17*, 396–408. doi:[10.1037/1040-3590.17.4.396](https://doi.org/10.1037/1040-3590.17.4.396).
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220–247. doi:[10.1207/s15327957pspr0803\\_1](https://doi.org/10.1207/s15327957pspr0803_1).
- Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Procedures and techniques for developing grounded theory*. Thousand Oaks, CA: Sage.
- Sylvester, B. D., Standage, M., Dowd, A. J., Martin, L. J., Sweet, S. N., & Beauchamp, M. R. (2014). Perceived variety, psychological needs satisfaction, and exercise-related well-being. *Psychology and Health, 29*, 1044–1061. doi:[10.1080/08870446.2014.907900](https://doi.org/10.1080/08870446.2014.907900).
- Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychological Assessment, 16*, 231–243. doi:[10.1037/1040-3590.16.3.231](https://doi.org/10.1037/1040-3590.16.3.231).
- Williams, D. M., & Rhodes, R. E. (2016). The confounded self-efficacy construct: Conceptual analysis and recommendations for future research. *Health Psychology Review, 10*, 113–128. doi:[10.1080/17437199.17432014.17941998](https://doi.org/10.1080/17437199.17432014.17941998).
- Williams, H. W. (2008). Characteristics that distinguish outstanding urban principals: Emotional intelligence, social intelligence and environmental adaptation. *Journal of Management Development, 27*, 36–54.
- Willis, G. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

# Chapter 3

## Contributions of Response Processes Analysis to the Validation of an Assessment of Higher Education Students' Competence in Business and Economics

Sebastian Brückner and James W. Pellegrino

### Increasing Importance of Response Processes Analysis in Validation Research

The *Standards for Educational and Psychological Testing* (2014) present various sources of evidence that can be used to evaluate a proposed interpretation of test scores for a particular purpose, such as evidence based on test content, internal structure, correlations with other variables, and consequences of testing. In contrast to those four types of evidence, response processes analysis has been given very little attention in the literature on validity (e.g., Leighton, 2004; Newton & Shaw, 2014; Pellegrino, Baxter, & Glaser, 1999; Zumbo & Chan, 2014). Response processes analysis has been largely ignored in the validation of test score interpretations related to learning constructs of significance in higher education (e.g., subject-specific knowledge in business and economics (B&E), mathematics, biology) in preference to psychological constructs such as intelligence and general problem solving skills (Ercikan et al., 2010; Leighton, 2013). This seems problematic considering that common validation evidence in achievement test development (e.g., analyzing the factorial structure of test scores or conducting interviews with lecturers and having them rate the items on how closely they reflect curricular standards) provides no explanation regarding possible misinterpretations of items or the underlying information processing operations associated with responding to items

---

S. Brückner (✉)

Department of Business and Economics Education, Johannes Gutenberg-University Mainz,  
Jakob-Welder-Weg 9, 55099 Mainz, Rheinland-Pfalz, Germany  
e-mail: [brueckner@uni-mainz.de](mailto:brueckner@uni-mainz.de)

J.W. Pellegrino

University of Illinois at Chicago, Learning Science Research Institute,  
1240 W. Harrison St. Suite, 60607-7019 Chicago, IL, USA  
e-mail: [pellejw@uic.edu](mailto:pellejw@uic.edu)

© Springer International Publishing AG 2017

B.D. Zumbo, A.M. Hubleby (eds.), *Understanding and Investigating Response Processes in Validation Research*, Social Indicators Research Series 69,  
DOI 10.1007/978-3-319-56129-5\_3

within the given domain. Such evidence refers to the output (e.g., test and item scores) or input (student-related features such as gender, socioeconomic status, and opportunities to learn) of an assessment.

Response processes analysis has been discussed with regard to its relationship to the psychometric quality of test items. Earlier research showed that individual items often could be interpreted by students in ways far more diverse than intended by the test developer or test user. This means that different mental operations could be responsible for item scores (Brückner & Pellegrino, 2016). Conversely, mental operations identical in terms of structure and content could lead to different item scores (Loevinger, 1957; Rulon, 1946; Turner & Fiske, 1968). Turner and Fiske (1968) analyzed mental operations qualitatively and then quantified the mental operations and correlated them with statistical evidence based on item scores (e.g., item discrimination). Subsequently, analysis of response processes established itself within the literature on information processing approaches to understanding performance on intelligence tests and aptitude tasks, wherein a need for the development of task performance models was recognized as part of the validation effort (e.g., Pellegrino & Glaser, 1979). Messick (1989) included such response processes analyses to substantiate his concept of a progressive validation matrix that focused on construct validity. Although the recent development of validity concepts underscores the significance of process analysis, the emphasis varies in terms of construct validity (Borsboom, Mellenbergh, & van Heerden, 2004; Lissitz & Samuelsen, 2007; Zumbo, 2009). In his demand for a revision of classic validation concepts, Zumbo (2009, p. 73) stresses the importance of explaining test scores: “One of the limitations of traditional quantitative test validation practices (e.g., factor-analytic methods, validity coefficients, and multitrait-multimethod approaches) is that they are descriptive rather than explanatory.” Lissitz and Samuelsen go even further, prioritizing the analysis of response processes relative to other analyses that refer to the nomological interrelations of different constructs. They maintain that analysis of response processes, along with evaluation of test content and reliability of a test, is an essential part of an internal test evaluation: “The area of the cognitive analysis of a test is one of the most productive and promising areas in psychometric application today” (Lissitz & Samuelsen, 2007, p. 445).

Kane (2004, 2013) argues that three fundamental types of test score interpretation should be taken into account when formulating an interpretive argument for subsequent validation studies: scoring, generalizing, and extrapolating. These kinds of interpretation are based on a logical structure of a validity argument (e.g., Angell, 1964; Toulmin, 1958) and can be used in the specification of the interpretive breadth of a response processes analysis. Kane (2013, p. 10) claims that the scoring inference “takes us from the observed performances to an observed score” and, according to Toulmin (1958), these performances are the data and the scores are the claim we wish to make. Thus, it is important to know what the performances on an item are in order to justify the scoring procedure. Because response processes analysis is a way to explain a test score (Zumbo, 2009), several authors assign response processes analysis to the level of scoring (Brückner & Kuhn, 2013; Howell, Phelps, Croft, Kirui, & Gitomer, 2013). Response processes analysis can also refer to generalizations based on justified and reasoned scoring procedures.

Thus, a generalization in the interpretation of test scores can be defined as the justification of aggregating scores on individual test items to reach an overall test score and goes beyond the single tasks and performance on them – “we typically generalize over the tasks included in the test or over test forms” (Kane, 2013, p. 19). In contrast to scoring inferences, here the scores are already defined and can be used additionally. This means that, in the test score interpretation, the mental operations are used to explain the meaning of a score on a task and that the mental operations are comparable to those taking place while completing other similar tasks. Extrapolation in interpreting test scores is when mental operations exhibited while responding to an item on a specific test are believed to occur when completing tasks on tests in general within a larger domain. For example, an operation relevant for solving algebraic problems also may be relevant for solving geometric problems or mathematics-related problems in general.

Recently, the analysis of response processes, especially the interpretation of their cognitive underpinnings (here referred to as mental or cognitive operations), has further established itself in research associated with arguments regarding the cognitive validity of an assessment (Baxter & Glaser, 1998; Pellegrino, DiBello, & Goldman, 2016; Shavelson, 2013). Thus, Messick’s (1989) approach of integrating response processes analysis into validation has been elaborated by emphasizing the significance of cognitive, construct-relevant operations and distinguishing such evidence from an exclusive focus on nomological networks. Cognitive operations are no longer understood merely as part of an overarching construct, as was the case in the classic model of construct validity. These processes now are the focus of the analysis, as they form the constitutive parts of the constructs. In more recent validity frameworks, emphasis has been placed on the significance of response processes analysis as part of the evidentiary argument for those aspects of validity that are especially important for instructionally relevant uses of assessment results (DiBello, Pellegrino, Gane, & Goldman, 2017; Pellegrino et al., 2016). In such contexts of assessment use, primary interest is which cognitive and non-cognitive operations (e.g., knowledge, motivation) the student has acquired as a consequence of the processes of instruction and learning, the evidence of which should therefore be elicited in the response behavior to the items on a test in a specific domain. As discussed below, it is both useful and necessary for response processes analysis to be connected to a validation framework that explicitly takes into account learning in a domain (see also DiBello et al., 2017; Pellegrino et al., 2016).

### **Situating Response Processes Analysis in an Instructional Validation Framework**

Pellegrino et al. (2016) developed a framework that builds on and reflects multiple and complementary facets of validity as articulated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014); however, they organized and prioritized them to evaluate the expected interpretive uses of assessments



intended to function in close proximity to the processes of teaching and learning. In this framework, three facets of validity are considered: cognitive, instructional, and inferential (Pellegrino et al., 2016).

*Cognitive Validity* is concerned with those aspects of knowledge and reasoning that students should use to solve items on a test. The latter reflects theories and models of the domain-specific construct and can be based on standards, curricula, workplace requirements, cognitive research, or other kinds of arguments and evidence that help define the construct that is being assessed and should be predominant in the interaction of students and items during the process of answering a question or solving a problem (Pellegrino et al., 2016).

*Instructional Validity* reflects alignment of the assessment tasks with instructionally relevant aspects of the domain (e.g., knowledge and skills targeted by curriculum materials, content standards, and/or empirical learning progressions, and that are the focus of instructional materials, methods, and activities). Moreover, an instructionally valid assessment should provide instructors with guidance on how to interpret and use test results adequately to guide educational decision making and practices (e.g., formative and summative assessment) (Pellegrino et al., 2016).

*Inferential Validity* implies interpreting and modeling the data (qualitatively and/or quantitatively) relative to inferences about the construct being assessed. While the first two facets of validity might reflect evidence relative to a qualitative or explanatory aspect of what the assessment is intended to do (Zumbo, 2009), evidence associated with inferential validity is related to accurate diagnosis, interpretation, and reporting of the “meaning” of a student’s performance. Models can come from multiple interpretive frameworks such as frequentist or Bayesian statistics, classical test theory, or item response theory, and they should give inferential insight into the fit of a student’s performance to the construct being measured (Pellegrino et al., 2016).

As argued by Pellegrino et al. (2016), all three facets of validity can be supported by evidence derived from response processes analysis and therefore can contribute to an overarching validity argument (Kane, 2013; AERA et al., 2014). The construct-relevant cognitive facets should become clear upon students’ interaction with the test items and should support a correct response, while construct-irrelevant cognitive facets should occur rather infrequently and be largely responsible for incorrect responses (see, for example, Gorin, 2006). Whether the scores are relevant for instruction depends on whether the knowledge and cognitive operations required for the assessment match the goals of the curriculum or educational standards and on their interpretability relative to instructional inferences (DiBello et al., 2017). In addition, the individual mental operations should provide explanations for item scores and therefore contribute to the findings – both in terms of content and construct-relevance – of statistical arguments (Zumbo, 2009). For selected response items, consideration is given to whether certain mental operations occur while responding to the item, and whether selecting a distractor indicates a significant misunderstanding that is meaningful for instruction (Luecht, 2007). Also, if the response processes data are available in a quantitative form, they can be related to dichotomous, trichotomous, or other partial credit scoring of the item response, and statistically analyzed.

According to Kane's (2004, 2013) validation concept, conducting response processes analysis requires theoretical specification of the interpretive argument before an evidence-based, empirical assessment of the warrants in support of that argument can be made within an overall validity argument. The three facets of validity discussed above underlie current thinking that an assessment is not valid or invalid *per se* (i.e., validity is not a property of the instrument), but rather validity has to be judged relative to the intended interpretive use of the results (Kane, 2013). According to several scholars (Baxter & Glaser, 1998; Borsboom et al., 2004; Brückner & Kuhn, 2013; Kane, 2013; Messick, 1989; Pellegrino et al., 2016; Pellegrino & Glaser, 1979), a response processes analysis concerned with the three aforementioned facets of validity should involve the following three steps.

1. First, the construct to be analyzed needs to be described and the construct-relevant and construct-irrelevant mental operations in terms of knowledge and reasoning (cognitive validity), as well as the meaning of the assessment relative to the curriculum for, and instruction in, the domain (instructional validity) need to be operationalized. This is much easier to accomplish if the domain is identified and the construct to be assessed described during a construct-centered and principled design process in which the assessment framework and critical operational variables are thoroughly defined (see e.g., Mislevy & Haertel, 2006; Pellegrino, Chudowsky, & Glaser, 2001; Wilson, 2005).
2. Second, a sampling strategy needs to be chosen, because in-depth response processes analysis cannot be conducted with as large a number of students or items as can be done in large empirical field studies focused primarily on item scores (Leighton, 2004). It is important that the students and items selected reflect important cognitive and instructional validation criteria as defined in step 1. For a response processes analysis, it is necessary that items and students selected differ in features considered relevant from a cognitive and instructional validity perspective. Furthermore, the sampling should support the inferences that the researcher desires to make. For example, to make statistical inferences (inferential validity), criteria important for generalizability (e.g., random samples) should be used, whereas to make qualitative comparisons, purposeful samples might be adequate (Creswell, 2014; Patton, 2015; Teddlie & Tashakkori, 2009). As another example, a higher level of academic performance might lead one to expect utterances of more construct-relevant cognitive operations aligned to academic performance. Extreme group comparison could be adequate in that case. In some studies, high and low test scores were used for grouping so that, afterwards, individual types of operations could be examined more closely in relation to differences in academic performance (e.g., Baxter & Glaser, 1998; Leighton, 2013; Pellegrino & Glaser, 1979; Thelk & Hoole, 2006). Doing this, category-related criteria (e.g., whether a respondent belonged to a specific professional group or had graduated) or ordinal and metric criteria (e.g., the grade upon graduation) could be used for the sampling. If one wishes to make an inferential claim (e.g., analyzing correlations), it is necessary to choose respondents at least partially randomly and according to certain criteria (e.g., students' progress in



their course of study measured by years of study). Thus, distributions in larger samples or from populations should be considered when selecting a sample for response processes analysis (e.g., Brückner & Pellegrino, 2016).

3. Third, an appropriate method for assessing typical kinds of operations must be chosen. For example, to investigate operations that tap attention and perception of the respondents (e.g., reading comprehension or translation skills), eye-tracking could be a suitable choice (e.g., Afflerbach & Cho, 2009; Gorin, 2006). If the focus is on constructs reflecting learning in a specific domain for which understanding of particular facts, concepts, or principles is important (e.g., social sciences), classic interviewing techniques or think-aloud interviews could be effective (e.g., Brückner & Kuhn, 2013; Chi, 1997; Leighton, 2004; Thelk & Hoole, 2006). Regardless, each approach ultimately involves a comparison of the empirically assessed mental operations to the claims made about the occurrence of processes with regard to cognitive, instructional, and inferential validity.

## **Response Processes Analysis in the Domain of B&E**

### ***Modeling Mental Operations***

In the domain of B&E, assessments often are developed on the basis of cognitively specified knowledge structures reflecting both declarative, procedural and other forms of knowledge. The former is defined as knowledge of or about an economic phenomenon and it relates to recall of economic concepts, facts, and principles. The latter is defined as knowledge that operates on such facts and principles and enables one to dynamically execute a specific mental operation, carry out a set of operations, or combine them using logical inferences, routines and associations (Arts, 2007). At a process level, the two forms of knowledge can be operationalized in such a way that declarative knowledge refers to significant allocation of the response to retrieval of, and discussion about, domain-specific knowledge (e.g., of the meaning of return on investment, SWOT analysis, productivity, or supply and demand) (Arts, 2007; Brückner, *in press*; Davies, 2006). Procedural knowledge goes beyond such recall and, at a process level, it points mainly to the dynamic and inferential use of knowledge structures (e.g., reasoning with the meaning of domain-specific concepts and applying them to the context of the problem) (Minnameier, 2013; Mislevy, 1994). The interaction of both components is then characterized by what Beck (1993) perceives as typical thinking for B&E in terms of “quantitatively optimized thinking” (p. 10). Thus, in B&E in higher education, it is possible to distinguish at a process level between mental operations that refer to the economic concepts and logical or deductive inferences. As stated by some authors (e.g., Arts, 2007; Brückner, *in press*), the elaboration of concepts can be correct or it can be associated with naïve imagination about economic concepts and, therefore,

incorrect. Nevertheless, in each case, this elaboration provides clarification of an economic fact or principle and, therefore, can be regarded as part of declarative economic knowledge. Deductive inferences are rather dynamic and therefore can be considered part of procedural economic knowledge (Arts, 2007; Minnameier, 2013). In what follows, we consider three cognitive operations with regard to their reflection of construct-relevance or construct-irrelevance in the B&E domain.

**Correct Elaborations on B&E Concepts** The purpose of having students elaborate on economic concepts is to determine whether they are able to explain and apply these concepts based on the expertise they have acquired over the course of their studies. According to Arts (2007, p. 19), these concepts can be perceived as a “class of phenomena” in B&E that helps to “reduce and characterize (managerial) phenomena into powerful and rather short labels.” They constitute the core of economic thinking and enable novices in the domain to access and acquire systematically knowledge in the domain (Davies, 2006). Some authors who prefer a purely inferential model also refer to the elaboration of said concepts as abductive inference (Minnameier, 2013; Mislevy, 1994). By comparing a student’s reasoning with explanations provided in multiple sources such as appropriate textbooks, surveys of professors and lecturers, or in analyses of lecture materials, it is possible to determine the degree of correctness of an explanation. Elaboration of B&E concepts can also be regarded as part of more complex strategies such as forward reasoning. Forward reasoning is an operation frequently used by B&E experts to solve B&E test items. However, it involves more than merely elaborating correctly on B&E concepts. It also encompasses identifying important aspects, setting goals, drawing accurate conclusions, and so on. (Brückner & Pellegrino, 2016).

**Deductive Inferences** Construct-relevant operations can be described using deductive inferences. Inferences generally can be defined as “transformations on literal information given in the original case text” (Arts, 2007, p. 19). A deductive inference can be understood as a logical conclusion derived from a combination of at least two statements (Minnameier, 2013; Mislevy, 1994). One statement, for instance, could be that the demand for a product is increasing. A second one could be that the supply curve remains constant. A deductive conclusion based on these two statements would be that the product’s price would have to be raised. Evidently, prior knowledge plays a great role in generating accurate deductive inferences when explaining economic phenomena. Deductive inferences, therefore, are an integral part of economic thinking (Arts, 2007; Brückner, *in press*; Minnameier, 2013). Thus, they also compose a central component of general economic problem solving operations such as forward reasoning and are relevant even for simple deductive thoughts such as summarizing linguistic information in a task statement, as is the case with paraphrasing (Brückner & Pellegrino, 2016).

**B&E Heuristics** Given high variability among beginning students in their prior education in B&E, many do not come to the domain with any detailed economic knowledge or experience. Therefore, they should be expected to succeed less often in adequately exploring a concept. Furthermore, some authors indicate that laymen

or novices may resort to general understanding or simplified conclusions (e.g., heuristics) and take on a polarizing perspective relative to the domain of B&E: “If the public tries to make sense of it [economics] nonetheless, it must impose some simpler structure or rely on heuristics” (Leiser & Aroch, 2009, p. 373). Leiser and Aroch (2009) describe how students often allocate an emotional charge to economic phenomena and perceive them as good or as part of vicious cycles. Domain experts also refer to the first case as the Good-begets-Good heuristic, which determines economic thinking. For instance, if a student claims that low costs are optimal, or low numbers are always good, they take on a simplified view which allows them at least rudimentary, though often erroneous, access to economic thinking. In contrast to the previously mentioned operation of correct elaboration, there should be less of a relationship of heuristic use to other operations relevant for the solution to a problem, such as forward reasoning, since economic heuristics reflect an unintended and mostly construct-irrelevant test taking behavior.

### ***WiwiKom Test Items to Assess Students’ Mental Operations Involved in Responding to the Items***

Test items used in B&E (e.g., from the Major Field Test (MFT), the Business Administration Knowledge Test (BAKT), or the Test of Understanding in College Economics (TUCE)) assess mostly the cognitive parts of students’ competence in B&E. They typically require more than factual knowledge and can be administered to assess higher education students’ reasoning and evaluation skills while solving problems in this domain (Bielinska-Kwapisz, Brown, & Semenik, 2012; Gröbner, Wilhelm, Wittmann, & Milling, 2002; Walstad, Watts, & Rebeck, 2007; Zlatkin-Troitschanskaia, Förster, Brückner, & Happ, 2014). To solve such problems, consideration must be given to the key concepts of B&E as well as to ways of applying them.

In our investigation of how response processes analysis contributes to obtaining evidence supporting possible claims about the cognitive, instructional, and inferential validity of B&E assessments, we examined higher education students’ performance on 19 standardized test items on B&E knowledge tests developed in the German project Modeling and Measuring Business and Economic Competencies of Students and Graduates (WiwiKom) (Zlatkin-Troitschanskaia et al., 2014). The items were multiple-choice and were adapted and translated from the international TUCE (Walstad et al., 2007) and Examina General para el Egreso de la Licenciatura Administracion (EGEL) (Vidal Uribe, 2013) to fit the German higher education context. Despite the validity claim that the WiwiKom Test assesses construct-relevant knowledge and operations, there still is limited evidence as to whether this is actually the case and what the implications might be for administration and use of the tests and scores in higher education instruction.

All items on the B&E knowledge tests administered in this study consisted of a situational item stem and four response options, one of which was correct (attractor); the other three were incorrect (distractors). The two formats of the items were

classic multiple-choice and complex multiple-choice (e.g., matching and sequencing items) (Zlatkin-Troitschanskaia et al., 2014).<sup>1</sup> The 19 items were considered representative of, and selected according to, theoretical criteria of content area and cognitive requirements from an item pool of 170 items that comprehensively assess the B&E competency of students in higher education (Brückner & Pellegrino, 2016) by covering important content areas and key concepts and forms of reasoning in that domain.

### *Student Sample*

The selection of students for such an analysis influences the robustness and utility of the response processes data relative to the validation of score meaning. The aim of this study was to show that the tasks developed in WiwiKom validly assess knowledge and mental operations acquired over the course of studies and therefore can provide insight into the cognitive, instructional, and inferential aspects of the validity of the test score interpretation (Kane, 2013). For this purpose, 20 students were selected in a purposeful random sampling (Creswell, 2014; Patton, 2015; Teddlie & Tashakkori, 2009); sampling was purposeful in that the students were selected from a larger sample (N = 882) according to the criterion of study progress (years of study) (Brückner & Pellegrino, 2016). This selection criterion was meant to account for the fact that students who were further along in their studies should have been able to respond to items more successfully and, in doing so, demonstrate the use of construct-relevant mental operations more frequently than students with less domain-relevant course experience (Arts, 2007). The findings from the WiwiKom project's field studies, where years of study have related positively to overall test scores, illustrate support for one part of this expectation for students in several countries (Förster et al., 2015). The question then is whether successful performance is, in fact, associated with application of relevant knowledge and reasoning operations in the B&E domain.

### *Think-Aloud Method*

Eye-tracking, verbal probing, and concept mapping are all suitable methods to assess empirically mental operations (Gorin, 2006). For the purposes of this study, the think-aloud method was considered the most suitable to obtain evidence of

---

<sup>1</sup>Matching items requires the respondent to combine economic facts, statements, and concepts with more or less accurate explanations. Sequencing items requires the respondent to causally or chronologically arrange economic statements, facts, or principles in a sequence (e.g., with the goal to conduct an economic analysis or optimization, for example, the steps to optimize a production sequence of a company) (for further information on this format, see Parkes and Zimmaro 2016).

greatest interest and relevance regarding response processes. Verbal data indicating the mental operations during the processing of tasks were collected in three phases.

In the first (concurrent) phase (CVP), the students worked on the items and thought aloud without being interrupted by the interviewer. Only when the students were silent for more than five seconds did the interviewer ask them to “keep talking” (Ericsson & Simon, 1993). All response options for a given task initially were covered with a sheet of paper so that the students first had to consider the item stem before evaluating the different response options. In the second phase, the students summarized what they were doing while solving the items. This was necessary because the concurrent verbal reports were sometimes unstructured and resembled a stream of consciousness. By summarizing the response processes, the interviewer obtained more structured insight into the various phases the student went through while responding to each item. In the retrospective or debriefing third phase, the interviewer was allowed to ask questions to clarify some fragmented utterances and to identify difficulties in the students’ response processes. In doing so, the students were led to reflect on some of their utterances from the concurrent phase. Thus, the data obtained from the retrospective phase were quite different as they reflected the students’ epistemic beliefs more clearly (Leighton, 2013). They supplemented the data obtained during the other two phases and provided details for comprehensive insight into the students’ response processes.

The data from the think-aloud interviews were transcribed and the transcriptions were coded to determine whether the operations that students were expected to demonstrate with reference to relevant concepts and inferences were indeed observed in the respective response situations. Because all 20 students responded to the 19 items, a total of 380 response processes protocols were analyzed. The transcription and coding were conducted using MAXQDA 11.

The following section presents selected results of the protocol analysis and discussion of the findings is organized in terms of their contribution to an argument about the various facets of validity in the domain of B&E relative to teaching and learning in this domain in higher education.

## Results

### *Cognitive Validity*

As discussed above, the verbal protocols were expected to contain evidence of the mental operations occurring while focusing on domain-relevant concepts and reasoning as well as those that are more domain-general such as logical reasoning in the sense of deductive inferences. Both types of concept-oriented operations were found in the data, abductively illustrating the meaning of individual concepts and thus generating a mental representation of the concept. Similar to the findings of previous studies (Arts, 2007; Brückner, 2013; Minnameier, 2013), there are operations that operate on an elaborate knowledge base and link concepts of B&E to a correct

**Table 3.1** Frequency of occurrence of mental operations

Mental operations	N	M	SD
Correct elaborations	66	.1737	.3793
Economic heuristics	72	.1895	.3924
Deductive inferences	223	.5868	.4930

**Table 3.2** Example of correct elaboration of B&E concepts

CVP/A
...
<u>well, productivity is output divided by input-</u> at least I think so. #00:09:54-9#
...

meaning. In one task, for example, test respondents calculated the productivity of shoe manufacturing. In 66 of 380 response processes (17.37%) (Table 3.1), the respondents managed to establish concepts correctly (Brückner, [in press](#)).

For instance, Respondent A (Table 3.2) correctly observed that productivity was determined by the ratio of output to input.

It was expected that, with increasing expertise in the domain of B&E, the amount of deductive inferences would increase as well because they are directly linked to elaboration on concepts (Arts, 2007; Brückner, [in press](#)). Following if-then rules, deductive inferences were the most common inferences found in the data (223; 58.68%) and indicated logical-associative thinking (see Table 3.1). Respondent C, for instance, tried to calculate the cost per unit and correctly elaborated that to do so he had to divide the total cost by the amount of units, and subsequently he made the calculation through deductive inference. This led him to the conclusion that 10 euros per pair of shoes should be the correct answer (Table 3.3). Compared to the operation mentioned earlier, there were significantly fewer indications of elaborating on economic concepts than using the aductively generated meaning of these concepts (“I think the costs divided by the number”) for further deductive reasoning (“This means I have 10 euros per..”).

In 18.95% (72) (Table 3.1) of the response processes analyzed, respondents exhibited naïve perceptions of economic phenomena, which also was common in Leiser and Aroch’s (2009) study. This means that students often underestimate the complexity of B&E concepts and transfer a simplified meaning from everyday life to these concepts. In accordance with the Good-begets-Good heuristic, some participants allocated an emotionally positive meaning to a concept instead of a descriptive, scientifically elaborated description and associated it with another positive economic meaning such as turnover (Table 3.4). This simplification, however, is not elaborated, as an increasing turnover leads to increased productivity only under certain circumstances, for instance, when selling and purchasing prices as well as the input remain constant. Thus, unlike the two operations of correct elaboration and deductive inferences, this operation can be regarded as a rather construct-irrelevant operation.

**Table 3.3** Example of deductive inferencing

---

**CVP/C:**

#00:08:27-9# (reading silently) #00:08:40-0#  
 Oh, calculate productivity? (groan)  
 #00:08:42-6# #00:08:46-7#  
 I think you have the costs.  
 How do you calculate that again?  
 I'd have to estimate that now uhm-  
 I think the costs divided by the number.  
This means I have 10 euros per uhm per pair  
 ...

---

**Table 3.4** Example of economic heuristic/oversimplification

---

**CVP/B**

...  
[productivity]  
you look at the company's turnover.  
 ...

---

Overall, expected cognitive operations were reflected in the data and, as expected, deductive operations occurred most frequently, whereas correct and heuristic elaborations occurred relatively rarely. Furthermore, as illustrated below, greater expertise in solving B&E items was differentially associated with correct elaborations, deductive inferences, and use of economic heuristics, each of which is relevant for responding correctly to the items. Such findings provide evidence supporting the cognitive validity of the assessment tasks, and this interpretation is further supported by evidence of their connection to, and differentiation from, other construct-relevant cognitive operations as well as to successful task solution (DiBello et al., 2017; Pellegrino et al., 2016). For example, in a previous study, forward reasoning was found to be important for solving these B&E tasks (Brückner & Pellegrino, 2016). This leads to the expectation that mental operations, such as correct elaboration on economic concepts, as well as greater ability to think logically and purposefully, should be related to these other operations and to successful solution of the assessment tasks. In contrast, simply repeating the task in one's own words, as is the case with paraphrasing, requires merely drawing conclusions and summarizing several aspects according to the respondent's logical perception of a task rather than correctly abducting domain-specific concepts (Brückner & Pellegrino, 2016). Evidence of the occurrence of this operation is not associated with successful performance on tasks.

To analyse correlations among these various operations, dichotomous scoring was done such that the use of a mental operation during the response processes was scored as 1 and absence of the operation was scored as 0 (e.g., correct elaboration = 1, no correct elaboration = 0). Because forward reasoning and paraphrasing also were coded dichotomously (Brückner & Pellegrino, 2016), a crosstab with four fields was created to examine the association between the occurrence of each pair of operations. As can be seen in Table 3.5, a  $\chi^2$  test of independence was conducted to analyze the association between forward reasoning and correct elaboration.



**Table 3.5** Correlation between forward reasoning and correct elaboration on B&E concepts

			Correct elaboration		
			Absent	Present	Total
Forward reasoning	Absent	n	145	15	160
		%	38.16	3.95	42.11
	Present	n	169	51	220
		%	44.47	13.42	57.89
Total	n	314	66	380	
	%	82.63	17.37	100.00	

$\chi^2(df) = 12.30(1); p < .001; \omega = .179$

Because in one cell the expected frequency was 3.95 and hence slightly below the required expected frequency of 5, an exact Fisher-Freeman-Halton test (e.g., Lydersen, Pradhan, Senchaudhuri, & Laake, 2007) was calculated in addition and affirmed the robustness of the results ( $p < .01$ )

**Table 3.6** Cohen’s  $\omega$  values for the relationship between the operations of forward reasoning and paraphrasing and evidence of correct elaboration on B&E concepts, economic heuristics, and deductive inferences (N = 380)

(Cohen’s $\omega$ )	Mental operations		
	Correct elaborations	Economic heuristics	Deductive inferences
Forward reasoning	.18***	-.12*	.15**
Paraphrasing	.00	-.04	.15**

*Note.* Here, only the effect sizes are marked with + for a positive correlation and – for a negative correlation. M represents the mean frequency of the occurrence of an operation  
\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$

Following a multitrait-multimethod approach (Campbell & Fiske, 1959), the effect sizes (Cohen’s  $\omega$ )<sup>2</sup> (Cohen, 1988) and their significance were determined for several such operation pairings. The results are summarized in Table 3.6. Correct elaboration, for instance, occurred in approximately 17% of 380 response processes [M(SD) = .17(.38)] (Table 3.1) and there was no independence from forward reasoning with an effect size of  $\omega = .18$ .

Although the effect sizes were modest, they support the conclusion that correct elaboration, economic heuristics, and deductive inferences are linked to other mental operations, and they illustrate which operations appear to be rather relevant or irrelevant to the construct. For instance, economic heuristics correlated negatively with forward reasoning, which was opposite to the correlations found for correct elaboration and inferential deductions. The latter are associated with more expert forms of reasoning and, as shown below, are associated with better scores on the tasks. Only deductive inferences showed a positive correlation to paraphrasing. This is not surprising, as repeating the text in one’s own words requires logical operations and simplification of sentence structures rather than elaboration and abductive operations (Minnameier, 2013; Mislevy, 1994).

<sup>2</sup>According to Cohen (1988), an effect with  $\omega = 0.1$  is classified as small effect,  $\omega = 0.3$  is a medium-sized effect, and  $\omega = 0.5$  is a large-sized effect.



**Table 3.7** Correlation between correct elaboration on B&E concepts and final response (Brückner, [in press](#))

		Final response			
			False	Correct	Total
Correct elaborations	Absent	n	166	148	314
		%	43.68	38.95	82.63
	Present	n	10	56	66
		%	2.63	14.74	17.37
Total	n	176	204	380	
	%	46.31	53.69	100.00	

$\chi^2(df) = 31.20 (1); p < .001; \omega = .287$

The question remains as to whether these construct-relevant cognitive operations are also associated with correct solution of the B&E tasks. Thus, correct item responses should correlate positively with construct-relevant facets such as correct elaboration and deductive inferences, and negatively or not at all with economic heuristics. Analogously to the WiwiKom project (Brückner, [in press](#); Zlatkin-Troitschanskaia et al., 2014), item responses were coded dichotomously, such that a correct response was coded with 1 and an incorrect response was coded with 0.

As Tables 3.7, 3.8, and 3.9 illustrate, the findings regarding the correlations of item responses with all three operation types are in accordance with expectations. Correct elaboration ( $\omega = .287$ ) and deductive inferences ( $\omega = .239$ ) correlated positively with correct responses, whereas correct responses correlated negatively with economic heuristics ( $\omega = - .292$ ). This is consistent with findings such as those of Arts (2007), showing that both drawing inferences and applying managerial concepts correlated positively with the accuracy of students' performance on test items. The infrequent occurrence of correct elaboration and economic heuristics can be attributed to the relatively large number of first-semester students in the sample (see the following section on instructional validity), who had not yet developed the necessary expertise to elaborate correctly on economic concepts. Of interest is the fact that the economic heuristics constitute only one part of construct-irrelevant incorrect operations and were used relatively infrequently. Thus, Brückner ([in press](#)) differentiates this type of operation from incorrect elaborations and "not knowing".

### *Instructional Validity*

To claim an assessment is instructionally valid, its connection to instruction and the curriculum must be shown, including how it can provide instructors with information that will support their teaching. It is, therefore, important that the mental operations elicited while responding to items are aligned with key assumptions of curriculum and instruction. Instructors in higher education must be able to identify variations in students' understanding of B&E phenomena and intervene

accordingly. Evidence of instructional validity could come from several sources. For example, the meaning of the mental operations evoked while solving items can be aligned with information associated with: (a) the content of textbooks, (b) instructional practices, (c) students’ level of progress in a course of study, and (d) instructors’ evaluations of the item content. These four sources of evidence are examined in more detail below.

First, analysis of frequently used textbooks (e.g., Mankiw (2012) and Krugman and Wells (2015) for economics; Wöhe and Döring (2013) for business) has revealed that correct elaboration on B&E concepts and logical and deductive reasoning about them is central for learning B&E at the higher education level and should be reflected in a higher probability of solving B&E items correctly. Evidence of such a relationship was shown in Tables 3.7, 3.8 and 3.9. In addition, the textbooks were found to provide the optimal definitions of, and elaboration on, the concepts, which allows test developers to judge the correctness of the elaboration of a concept in a certain context (Tables 3.2, 3.3, and 3.4) and helps lecturers identify the instructional potential of tapping into the mental operations triggered by various tasks. For instance, the textbooks contained information on the concept of productivity, which allows Respondent A’s elaboration to be recognized as correct, a desirable outcome for a student respondent after completing a module from the B&E course. Thus, teaching students how to elaborate on B&E concepts correctly, providing them with definitions and principles of, and facts about, B&E, and initiating tasks to give students the opportunity to elicit various mental operations that foster learning in that domain.

**Table 3.8** Correlation between economic heuristics and final response

			Final response		Total
			False	Correct	
Economic heuristics	Absent	n	121	187	308
		%	31.84	49.21	81.05
	Present	n	55	17	72
		%	14.47	4.47	18.95
Total	n	176	204	380	
	%	46.31	53.69	100.00	

$\chi^2(df) = 32.311 (1); p < .001; \omega = - .292$

**Table 3.9** Correlation between deductive inferences and final response (Brückner, [in press](#))

			Final response		Total
			False	Correct	
Deductive inferences	Absent	n	95	62	157
		%	25.00	16.32	41.32
	Present	n	81	142	223
		%	21.32	37.37	58.68
Total	n	176	204	380	
	%	46.31	53.69	100.00	

$\chi^2(df) = 21.68(1); p < .001; \omega = .239$

Second, greater instructional potential comes from the operation of economic heuristics, which students employ but usually indicate flawed understanding. For instance, by associating productivity with turnover, Respondent B (Table 3.4) introduces a nominal value that is not part of the concept of productivity. However, this seems to make understanding the task easier for him, as this value is familiar to him from his everyday life and enables him to begin elaborating on concepts. In the further course of the module, the lecturer should respond by introducing the difference between real and nominal values, or between quantity and value in an enterprise. Thus, students' use of economic heuristics, reflecting naïveté and sometimes misunderstanding, is a starting point for lecturers to provide important instructional feedback, which necessitates developing good scaffolding skills (e.g., Hattie & Timperley, 2007; Puntambekar & Hubscher, 2005). Such misunderstanding is taken into account accordingly in the distractors. For instance, at least one distractor explicitly refers to an incorrect nominal view of productivity. The item then provides important diagnostic information that supports instructional validity. In previous studies of financial accounting (Vernooij, 2000) and macroeconomics (Leiser & Aroch, 2009), typical misunderstandings regarding B&E concepts were investigated and similar misunderstandings were elicited by the items administered in this study.

Third, examining instructional validity requires relating the verbalization of operations with the respondents' study progress. Students with more years of study were expected to be able to verbalize more frequently correct elaboration on economic concepts and deductive inferences and conversely rely less frequently on naïve concepts and economic heuristics. This can be most easily determined by use of exploratory correlation analyses. Similar to Arts' (2007) findings, in the present study, expertise (measured by study progress in the form of years of study) showed a positive correlation with correct responses as well as with the construct-relevant operations (Table 3.10). A rank correlation was used because years of study was a metric variable reflecting the fact that the students in this study were mostly beginners in their second year of study ( $M = 2.35$ ;  $SD = 1.49$ ) (Brückner & Pellegrino, 2016), whereas the final item responses and mental operations were scored dichotomously.

In addition to analyzing the correlations with study progress, evaluating the instructional relevance of the content of the items by professors and lecturers and its alignment with mental operations also is important in evaluating the instructional validity of an assessment. In an online survey, between three and eight professors and lecturers from different universities assessed the individual items on a seven-point Likert-type response scale (1 = very low to 7 = very high), answering, among others, a question about the extent to which the items covered content relevant to the curriculum. Overall, the items were assessed positively with regard to their relevance ( $M = 5.47$ ,  $SD = 0.85$ ) (Brückner, *in press*).<sup>3</sup> Item response accuracy correlated negatively with the estimated relevance of the content (Table 3.11) (Brückner, *in press*). This can be attributed to the likelihood that a large number of first-semester

---

<sup>3</sup>Thee of the items used were not included in the online rating, as they were newly developed in cooperation with experts so as to ensure that they were representative of the curriculum.

**Table 3.10** Rank correlations between years of study and each of accuracy of final response and mental operations

		Mental operations		
<i>Rho</i> (N = 380)	Final response	Correct elaborations	Economic heuristics	Deductive inferences
Years of study	.18***	.17***	-.12*	.23***

Note. \* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$

**Table 3.11** Rank correlations between content relevance and each of final response and mental operations

		Mental operations		
<i>Rho</i> (N = 320)	Final response	Correct elaborations	Economic heuristics	Deductive inferences
Content relevance	-.11*	.21***	-.16**	-.03

Note. \* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$

students had not yet acquired insight about economic content and therefore less frequently responded correctly to the content-relevant items. Furthermore, in accordance with expectations, the use of construct-relevant elaboration on economic concepts correlated positively with the content-based item evaluations, as it matched the content taught in the course and was reflected in the tasks. Similarly, a naïve understanding of concepts correlated negatively with relevant content. Surprisingly, while the frequency of deductive inferences appeared to increase with years of study (Table 3.10), use of this operation did not correspond with relevant content. This can be attributed to the rather limited construct relevance of deductive inferences, as illustrated above. However, it could be due to the fact that deductive inferences were sufficiently represented in the items but not in the curriculum, and/or that experts did not take deductive inferences into account upon rating the tasks and attended instead to the correct elaboration on economic concepts.

Overall, this analysis of empirically observed mental operations in relation to the content of textbooks, their instructional potential, the participants' years of study and expert ratings provide significant preliminary indications of instructional validity: the operations identified are important for cognition and for describing and evaluating learning B&E in higher education. However, this study has been merely a first step in a more comprehensive validity argument which needs to describe and identify mental operations important for learning in B&E. Further analysis could involve, for example, expert ratings by B&E professors and lecturers who could evaluate the individual items according to the importance of the individual operations and compare this assessment to the operations actually verbalized by respondents. Thus, a direct link could be made between the evidence related to cognitive validity and its bearing on conclusions about instructional validity. Meanwhile, the curricular estimates can provide some preliminary evidence regarding the curricular significance of the operations. Thus, the relationship between evidence of instructional validity and whether that same evidence is linked to inferential validity is only partially established in the current work. The convergent and discriminant

findings, however, confirm the cognitive and instructional relevance of the operations assessed in the tasks, supporting the overarching validity argument that construct-relevant operations contribute to the successful solution of economic tasks in higher education. The present analysis did not examine the extent to which operations correlated with the choice of individual distractors. Such evidence could add to the arguments for all three facets of validity of these assessments.

## Discussion

Response processes analysis is a part of assessment validation that has thus far been given relatively little attention compared to other methods of generating validity evidence. There appear to be multiple reasons that response processes analysis has been largely neglected. For example, Leighton (2004) observed several years ago that task performance models providing deeper insight into mental operations involved in responding to various test items are not applied very frequently in learning domains, despite calls for doing so that have appeared for quite some time (e.g., Pellegrino & Glaser, 1979). In comparison, general construct definitions and assessment frameworks that are more related to the overall test construct than to individual mental operations are naturally a core element of validation. This is why, unfortunately, in many higher education domains little is known about the operations that occur during task solution. Another reason may be that the studies in which these operations are examined often are very complex in terms of the behavioral data obtained as well as the methods for analysis of those data. Think-aloud studies, eye-tracking, or cognitive neuroscience methods such as EEG, fMRI, and so on require technical equipment and training in the methods of cognitive psychological research (Gorin, 2006; Leighton, 2004). In addition, findings from such studies cannot be expected to be highly generalizable because the number of respondents participating in such studies often has to be kept small, partially due to time factors as well as the size of the resulting data corpus. In many cases, the publication of such research is often problematic unless the results are part of a larger validation effort with multiple data sources.

To address these difficulties, it is advisable to embed response processes analysis in more comprehensive validation designs so as to create a wider frame of reference for the studies and to increase the generalizability of the results and conclusions. In conducting response processes analysis in the validation of assessments, multiple frameworks should be considered to structure the work, such as the instructionally relevant assessment validation framework of Pellegrino et al. (2016), as well as Kane's (2013) three types of interpretations (i.e., scoring, generalization, and extrapolation). For example, clarification as to whether response processes analysis is used mostly for scoring, generalization, or extrapolation can be useful in order to prevent false expectations and to underline the significance of response processes analysis in validation.

Particularly in the domain of B&E, and especially in the context of higher education, too little is known about the knowledge structures and cognitive operations associated with the targets of learning and instruction which are supposed to be assessed using standardized tests. Towards that end, the current results have been discussed in the context of the validation framework proposed by Pellegrino et al. (2016) for instructionally relevant assessments. The results presented provide indications that evidence of cognitively and instructionally relevant mental operations can be derived from response processes data and that such operations are related to successful outcomes on the assessment tasks. Undoubtedly, analyses of the type presented in this paper should be complemented by analyses of additional forms of data such as response times, responses on questionnaires assessing metacognitive constructs or academic self-concept, or performance on intelligence tests, and they should be further analyzed using more accurate and complex methods such as multi-level models in which responses are clustered according to students (Brückner & Pellegrino, 2016). However, the priority of this study was to determine whether it was possible to use response processes results in a validation analysis that could be meaningful and valuable for higher education instruction and to present the results in the context of a validation approach that simultaneously considers three facets of validity – cognitive, instructional, and inferential. The results show how a validity argument might be constructed for instructionally supportive assessment in this domain. Furthermore, it is argued that findings derived from the analysis of response processes data can be meaningfully interpreted only if multiple sources of evidence are taken into account (e.g., expert interviews) and interconnected so that instructors and students can be supported in interpretations of the meaning of an assessment score relative to their use in the classroom. Given this frame of reference, assessments can be more accurately designed and integrated with materials for teaching and learning so as to support a process of formative assessment that provides direct feedback to the students about the progress of their learning. In doing so, instruction and learning in B&E in higher education can be improved.

## References

- Afflerbach, P., & Cho, B.-Y. (2009). Identifying and describing constructively responsive comprehension strategies in new and traditional forms of reading. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 69–90). New York, NY: Routledge.
- 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 Psychological Association.
- Angell, R. B. (1964). *Reasoning and logic*. New York, NY: Appleton-Century-Croft.
- Arts, J. (2007). *Developing managerial expertise: Studies on managerial cognition and the implications for management education*. Maastricht, Netherlands: University Library.
- Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 17, 37–45.
- Beck, K. (1993). *Dimensionen der ökonomischen Bildung. Meßinstrumente und Befunde. Abschlußbericht zum DFG-Projekt: Wirtschaftskundliche Bildung-Test (WBT). Normierung*

- und internationaler Vergleich* [Dimensions of economics literacy. Measuring instruments and findings. Closing report of the DFG project: Test of economic literacy (TEL). Standardization and international comparison]. Nürnberg, Germany: Universität Erlangen-Nürnberg.
- Bielinska-Kwapisz, A., Brown, F. W., & Semenik, R. (2012). Interpreting standardized assessment test scores and setting performance goals in the context of student characteristics: The case of the major field test in business. *Journal of Education for Business*, 87, 7–13.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- Brückner, S. (2013). Construct-irrelevant mental processes in university students' responding to business and economic test items: Using symmetry based on verbal reports to establish the validity of test score interpretations. *Brunswik Society Newsletter*, 28, 16–20.
- Brückner, S. (in press). *Prozessbezogene Validierung anhand von mentalen Operationen bei der Bearbeitung wirtschaftswissenschaftlicher Testaufgaben* [Process-related validation using mental operations during solving business and economics test items] (Doctoral dissertation). Johannes Gutenberg-Universität, Mainz. Landau, Germany: Verlag Empirische Pädagogik.
- Brückner, S., & Kuhn, C. (2013). Die Methode des lauten Denkens und ihre Rolle für die Testentwicklung und Validierung [The think-aloud method and its significance in test development and validation] In O. Zlatkin-Troitschanskaia, R. Nickolaus, & K. Beck (Eds.), *Lehrerbildung auf dem Prüfstand (Sonderheft). Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften* [Teacher education under scrutiny (special issue). Modeling and measuring students' competencies in business, economics and engineering] (pp. 26–48). Landau, Germany: Verlag Empirische Pädagogik.
- Brückner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multi-level models to validate an assessment of higher education students' competency in business and economics. *Journal of Educational Measurement*, 53, 293–312.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, 6, 271–315.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: SAGE Publications.
- Davies, P. (2006). Threshold concepts. How can we recognise them? In J. Meyer & R. Land (Eds.), *Overcoming barriers to student understanding. Threshold concepts and troublesome knowledge* (pp. 70–84). London, UK: Routledge.
- DiBello, L. V., Pellegrino, J. W., Gane, B. D., & Goldman, S. R. (2017). The contribution of student response processes to validity analyses for instructionally supportive assessments. In K. W. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning in the next generation of assessments. The use of response processes* (pp. 85–99). London, UK: Routledge.
- Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, 29, 24–35.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed). A Bradford book. Cambridge, MA: MIT Press.
- Förster, M., Zlatkin-Troitschanskaia, O., Brückner, S., Happ, R., Hambleton, R. K., Walstad, W. B., et al. (2015). Validating test score interpretations by cross-national comparison: Comparing the results of students from Japan and Germany on an American test of economic knowledge in higher education. *Zeitschrift für Psychologie*, 223, 14–23.
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25, 21–35.
- Größler, A., Wilhelm, O., Wittmann, W. W., & Milling, P. M. (2002). *Measuring business knowledge for personnel selection in small and medium sized companies: Abschlussbericht zum*



- Projekt: Die Erfassung von Wirtschaftswissen zur Personalauswahl in KMU* (No. 44). Mannheim, Germany: Institut für Mittelstandsforschung der Universität Mannheim.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81–112.
- Howell, H., Phelps, G., Croft, A. J., Kirui, D., & Gitomer, D. (2013). *Cognitive interviews as a tool for investigating the validity of content knowledge for teaching assessments* (ETS Research Report No. RR-13-19). Princeton, NJ: ETS.
- Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, 2, 135–170.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.
- Krugman, P. R., & Wells, R. (2015). *Economics* (4th ed.). New York, NY: W.H. Freeman & Co Ltd..
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6–15.
- Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education*, 26, 136–157.
- Leiser, D., & Aroch, R. (2009). Lay understanding of macroeconomic causation: The good-begets-good heuristic. *Applied Psychology*, 58, 370–384.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437–448.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.
- Luecht, R. M. (2007). Using information from multiple-choice distractors to enhance cognitive-diagnostic score reporting. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education. Theory and applications* (pp. 319–340). Cambridge, UK: Cambridge University Press.
- Lydersen, S., Pradhan, V., Senchaudhuri, P., & Laake, P. (2007). Choice of test for association in small sample unordered r x c tables. *Statistics in Medicine*, 26, 4328–4343.
- Mankiw, N. G. (2012). *Principles of economics* (6th ed.). Mason, OH: South-Western Cengage Learning.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan Publishing.
- Minnameier, G. (2013). The inferential construction of knowledge in the domain of business and economics. In K. Beck & O. Zlatkin-Troitschanskaia (Eds.), *Professional and VET learning: Vol. 2. From diagnostics to learning success. Proceedings in vocational education and training* (pp. 141–156). Rotterdam, Netherlands: Sense Publishers.
- Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 439–483.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25, 6–20.
- Newton, P. E., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Los Angeles, CA: SAGE.
- Parkes, J., & Zimmaro, D. (2016). *Learning and assessing with multiple choice questions in college classrooms*. New York, NY: Routledge.
- Patton, M. Q. (2015). *Qualitative research & evaluation methods* (4th ed.). Thousand Oaks, CA: SAGE.
- 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., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
- 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.
- Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence, 3*, 187–215.
- Puntambekar, S., & Hubscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist, 40*, 1–12.
- Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16*, 290–296.
- Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist, 48*, 73–86.
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Los Angeles, CA: SAGE.
- Thelk, A. D., & Hoole, E. R. (2006). What are you thinking? Postsecondary student think-alouds of scientific and quantitative reasoning items. *The Journal of General Education, 55*, 17–39.
- Toulmin, S. E. (1958). *The uses of argument* (updated ed.). Cambridge, UK: Cambridge University Press.
- Turner, C., & Fiske, D. W. (1968). Item quality and appropriateness of response processes. *Educational and Psychological Measurement, 28*, 297–315.
- Vernooij, A. (2000). Tracking down the knowledge structure of students. In L. Borghans, W. H. Gijsselaers, R. G. Milter, & J. E. Stinson (Eds.), *Educational innovation in economics and business V. Business education for the changing workplace* (pp. 437–450). Dordrecht, Netherlands: Kluwer Academic Publishers.
- Vidal Uribe, R. (2013). Measurement of learning outcomes in higher education: The case of general in Mexico. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education* (pp. 137–146). Rotterdam, Netherlands: Sense Publishers.
- Walstad, W. B., Watts, M., & Rebeck, K. (2007). *Test of understanding in college economics: Examiner's manual* (4th ed.). New York, NY: National Council on Economic Education.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.
- Wöhe, G., & Döring, U. (2013). *Einführung in die allgemeine Betriebswirtschaftslehre* (25., überarb. und aktualisierte Aufl.) [Introduction into general business administration] (25th rev. vol.). München, Germany: Vahlen.
- Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from the German assessment of business and economics competence. In H. Coates (Ed.), *Assessing learning outcomes: Perspectives for quality improvement* (pp. 175–197). Frankfurt am Main, Germany: Lang.
- 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, NC: Information Age Pub.
- Zumbo, B. D., & Chan, E. K. (2014). *Validity and validation in social, behavioral, and health sciences* (Vol. 54). New York, NY: Springer International Publishing.

# Chapter 4

## Ecological Framework of Item Responding as Validity Evidence: An Application of Multilevel DIF Modeling Using PISA Data

Michelle Y. Chen and Bruno D. Zumbo

Researchers and decision makers in education have become increasingly interested in results from international assessments. To draw valid inferences about student academic achievement from such data, many factors need to be taken into account. Zumbo and colleagues (Zumbo & Gelin, 2005; Zumbo et al., 2015) have suggested that to truly understand the item responses, different explanatory sources, such as psychological and cognitive factors, physical and structural settings of the community, as well as the social context need to be explored. This ecological view of item response or test performance rests on an evolutionary, adaptive view of human beings in continuous interaction with their environment, with particular consideration for measurement validity and response processes. Viewed in an ecological framework, item responses and test performance cannot be simply attributed to the individuals or the environment, but to the relationship between the two.

As emphasized in some developmental psychology theories (e.g., Bronfenbrenner, 1979), the ecological conditions of individuals fosters their psychological growth. Ecological conditions described in developmental psychology theories usually include the environments at home, school, and the workplace. Building on such theories, Zumbo and colleagues (2015) described the ecology of item responding with the item responding embedded in a multiplicity of contexts (see Fig. 4.1). Views of measurement validity by Messick, Zumbo, and others focus on evidence about why and how people respond as central evidence for measurement validation. In line with Messick's (1989, 1995) articulation of substantive validity, the ecological model of item responding provides a contextualized and embedded view of

---

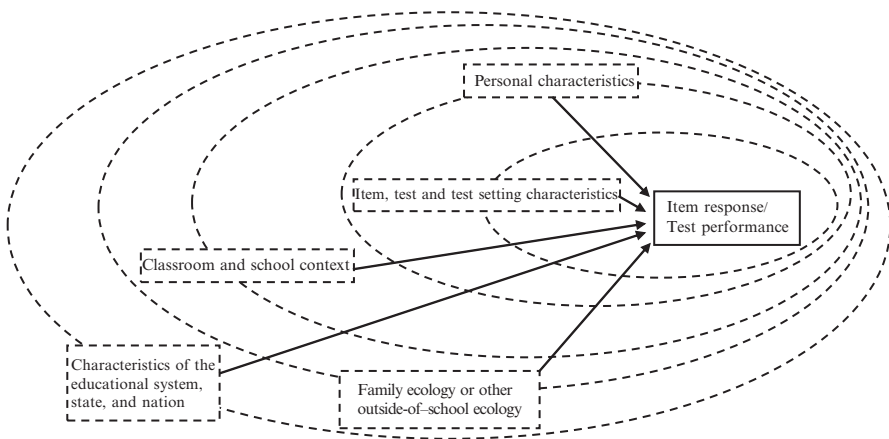
M.Y. Chen • B.D. Zumbo (✉)

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: [michellec2004@gmail.com](mailto:michellec2004@gmail.com); [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca)

response processes conceptualized as a type of situated cognitive framework for test validation (Zumbo, 2009; Zumbo et al., 2015).

The ecological model describes the enabling conditions for the abductive explanation for variation in test performance (Stone & Zumbo, 2016; Zumbo, 2007, 2009). As such, the study of response processes is guided by a contextualized pragmatic form of abductive explanation. In terms of the process of validation (as opposed to validity, itself), the methods described herein work to establish and support the inference to the best explanation – i.e., validity itself; so that validity is the contextualized explanation via the variables offered in the ecological model, whereas the process of validation involves the myriad methods of psychometric and statistical modeling (Zumbo, 2007). Zumbo’s abductive approach to validation seeks the enabling conditions via the ecological model through which a claim about a person’s ability from test performance makes sense (Stone & Zumbo, 2016; Zumbo, 2007, 2009).

Using educational testing as an example, five layers of an ecology of item responding and test performance are depicted in Fig. 4.1. In this case, the five layers are: (a) test and test setting characteristics; (b) personal characteristics; (c) classroom and school context; (d) family ecology or other outside-of-school ecology; and (e) characteristics of the education system and the nation state. Item and test properties, such as the content of the test, the format of the test, and the psychometric properties of that test, are at the lowest level in this ecological model. In the next layer are student characteristics, which may include biological sex, gender identity and expression, grade level, behaviors, attitudes, and other psycho-social characteristics of a student. Item properties and individual characteristics are the immediate context of item responding and thus test performance. Item and test characteristics reflect the properties of tasks that test takers or respondents need to engage in, and personal characteristics reflect the individuals engaging in the response process.



**Fig. 4.1** Ecological framework for item responding and test performance

These two layers represent the parties directly involved in the item responding process. At a further layer are classroom- and school-related factors, such as student-teacher relationship, peer behaviors, school type, and so on. These factors may, to a large extent, determine students' learning experience and available resources at school. The next layer represents outside-of-school factors. When students enter school, their knowledge, skills, and ability levels vary widely by their outside-of-school conditions such as their family's socioeconomic status, immigrant status, and their parents' parenting style. These out-of-school factors may continue to contribute to student test performances as the students progress through school. The last layer represents neighborhood, jurisdiction, and nation state context in which student characteristics, school contexts, and outside-of-school context are embedded. All of these contextual factors work together to influence the item responding process, and they may have a direct or indirect relationship with students' item response and test performance.

One of the central features of this ecological framework is that it explicitly illustrates the complexity of the ecology of the item responding. This ecological framework is proposed with aims to motivate a focus on contextual factors and to guide the development of contextual models to explain item responding via these enabling conditions guiding the abductive explanation. Without a conceptual framework organizing various aspects of the ecology of item responding, it is difficult to study the sources of item response or test performance variability systematically. By explicitly listing the contextual factors, researchers can easily notice the important factors or interactions between different factors that are left out in a study. Studies that have only focused on variables drawing from one or two aspects of an ecology system may either over- or under-emphasize the contribution of those particular aspects of the ecological system to item responding. This ecological approach intentionally expands the sphere of variables in a typical analysis of response processes.

To include various aspects of the ecology framework within a single set of analyses, appropriate modeling strategies are needed. The multilayer nature of item responding ecology fits well with the idea of multilevel modeling via mixture models. The key statistical concern with data collected from a multilevel system is the violation of the statistical independence assumption. Within a hierarchical system, lower-level observations are nested within a higher level factor. This nesting nature of observations is likely to produce some degree of similarity among the observations nested within the same unit, and thus these observations are not entirely independent from each other. For example, students from the same school may share some common views because their experiences at school are more likely to be similar compared to the experiences of their peers' from another school. When such non-independence is present and ignored in the data analysis, it shrinks the estimated standard errors and increases Type I error rates of the hypothesis testing. To avoid such issues, multilevel modeling or multilevel regression model, which distinguishes multiple levels of information in a model, should be used. Multilevel models estimate and incorporate the non-independence between observations directly into the analysis, so that cross-level effects are correctly estimated and tested. For

detailed overviews of multilevel modeling, please refer to Raudenbush and Bryk (2002), and Hox (2002).

As presented in the ecological framework of item responding, each layer of this framework (e.g., individuals, schools, community) can be an important unit of analysis. Researchers can build models that capture the layered structure of the item responding ecology, and determine how layers interact and impact a dependent variable of interest. It should be noted that, although the ecological framework is presented as a concentric circles model in Fig. 4.1, we do not assume a strictly nested structure among different layers in this framework. Instead, the relationships among some of these layers can be networked where different layers can have overlap and interaction. The figurative presentation of this framework only attempts to give an overall description of different aspects in the ecology surrounding an item response or test performance. The structure of this ecology framework does not directly correspond to the structure of a multilevel model. For example, personal characteristics and family ecology are distinguished as two different layers in the ecological framework of item responding and test performance. This does not imply that variables drawn from personal characteristics and variables drawn from family ecology should always be modelled at different levels in a multilevel regression model. To decide the structure of a multilevel model, the level of the variable being measured, the structure of the data, and the theory to be tested all need to be considered.

The purpose of this chapter is to link the conceptual framework of item responding ecology with a widely used modeling strategy. By doing so, we aim to motivate the application of this ecological framework in guiding the investigation of contextual factors in item responding process and test score validation. The usefulness of this approach is demonstrated using measurements of reading attitude in the Program for International Student Assessment (PISA) 2009 assessment. PISA is a large scale international assessment administered in more than 60 countries all over the world (OECD, 2010). It is a widely used data set to evaluate and understand student academic performance on reading, math and science literacy. Our focus is on how to think about and analyze background information (i.e., these enabling conditions) to explain item responding. We present a study to demonstrate how the item responses can be investigated to inform an ecological perspective. This study investigated item level measurement invariance between gender groups of a reading attitude scale. Multilevel regression methods to detect differential item functioning (DIF) are introduced and discussed. Understanding why (matched) groups perform differently on items provides a unique window into item response processes, and test validation. In line with the AERA/APA/NCME *Test Standards*' (2014) description of evidence based on response processes, the ecological model can contribute to answering questions about differences in meaning or interpretation of test scores across relevant subgroups and the extent to which capabilities ancillary to the phenomenon of interest may be differentially influencing test performance.

## Differential Item Functioning

DIF occurs when respondents from different groups show a differing probability of endorsement for an item after matching them on the ability or construct that the item is intended to measure (e.g., Zumbo, 1999, 2008). DIF studies are conducted to serve different purposes. For example, running DIF analysis is a common practice in standardized testing to address potential fairness issues. This practice is largely driven by policy and legislation. Testing DIF is also a way to rule out a lack of measurement invariance as an alternative explanation for observed group difference. In this example, however, DIF analysis is used as a method to explore the psychosocial processes of item responding by investigating whether contextual factors can explain the variability in the item responding across known groups.

Seeking the sources of DIF is a feature of the third generation of DIF (Zumbo, 2007). Most of the existent DIF studies focused on the investigation of item format, test content, and individual characteristics as the sources of DIF. The role of other more remote contextual factors, such as classroom and school characteristics, and jurisdictional and national differences, are usually left out in the investigation of DIF sources. As discussed in the ecological framework of item responding, sources of DIF can be very diverse and can include a lot more contextual factors beyond item properties and respondents' characteristics. Multilevel models are potentially useful for explaining DIF especially when the explanatory variables are at a higher level (e.g., group-level). These novel DIF methods provide us with a toolkit to build and test working hypotheses about why (matched) groups perform differently on items; hence providing, as we noted earlier, a unique window into item response processes, and test validation.

Guided by the ecological framework of item responding, we used multilevel logistic regression models to demonstrate an investigation of national-level indices in the explanation of gender DIF effect. Investigation of gender-related DIF is a common topic in the DIF literature. Gender is usually conceptualized as binary and used to separate individuals into different groups in such studies. When items are identified as showing DIF between different gender groups, item properties are often examined by content experts to identify possible sources of DIF. This practice has led to a focus on item properties, such as item format and item content in the explanation of DIF. The conceptualization of gender has been pushed from biological differences to a socially constructed concept. This shift in the definition of gender emphasizes the importance for us to set the gender DIF investigation in the social context. In other words, the sources of gender DIF should not be solely considered as certain item properties. The psychosocial environment in which the individuals live can contribute to the gender DIF effect as well.

Multilevel logistic regression models have been used to detect DIF in several studies (e.g., Balluerka, Gorostiaga, Gómez-Benito, & Hidalgo, 2010; Balluerka, Plewis, Gorostiaga, & Padilla, 2014; Swanson, Clauser, Case, Nungester, & Featherman, 2002). Swanson and colleagues (2002) proposed a two-level logistic regression model in which person responses are nested within items to detect DIF and explain its causes. In the models they used, the level-1 models (individual level)

are logistic regression models similar to those proposed by Swaminathan and Rogers (1990). The level-2 models are at the item level where the regression coefficients from level-1 models are treated as random variables. Swanson and colleagues' (2002) simulation study suggested the potential value of using a multilevel model to investigate DIF and sources of DIF. Balluerka and colleagues (2010, 2014) applied the multilevel model proposed by Swanson and colleagues (2002) to case studies and demonstrated the usefulness of that model in the investigation of uniform DIF. Van den Noortgate, De Boeck, and Meulders (2003) proposed a slightly different multilevel model to investigate DIF. In a later study, Van den Noortgate and De Boeck (2005) presented examples of using such mixed-effect logistic regression models in DIF investigation. In their examples, Van den Noortgate and De Boeck presented the possibility of treating item effects as random, treating group effects as random, or treating both item effects and group effects as random. Three-level logistic regression models were used in their examples. In their three-level models, item responses are at level-1, respondents are at level-2, and the highest level is the grouping of respondents (e.g., school). One of the common features shared by these models is that items or a group of 'similar' items are modelled together in one model. When modelled in this way, the DIF parameter for an item needs to be interpreted by contrasting it with the mean of the other DIF parameters.

As our focus in this study is to provide a method for connecting potential sources of DIF to contextual factors outside of test settings, we were interested in the potential of using national indices to explain the variability of DIF effects across nations. Instead of modeling all of the items together, we followed the traditional logistic regression methods for DIF investigation (e.g., Swaminathan & Rogers, 1990; Zumbo, 1999, 2008) and tested items for DIF one at a time. The multilevel models used in this study are a natural extension of the traditional logistic regression methods in which the level-1 models are the same as the logistic regression models described by Swaminathan and Rogers (1990) and Zumbo (1999, 2008). The level-2 models are at the national level, and national indices are added in later to explain the variability in the DIF effect across nations. The models used in this study test DIF for each item individually and the DIF effect tested is an average effect across level-2 units (i.e., nations, in this example). More details about how the multilevel logistic regression models are specified in our study are presented in the data analysis section.

Theoretically, both uniform DIF and nonuniform DIF can be investigated using multilevel logistic regression methods as proposed in our study. Only uniform DIF, however, is tested and attempted to be explained by national indices in this study. This is because the primary goals of this study are to (a) demonstrate an example of using multilevel models to investigate DIF, and (b) draw attention to contextual factors, such as social development factors, as potential sources of DIF effects. Adding explanatory variables to the nonuniform DIF effect results in a 3-way cross-level interaction, which can be difficult to interpret. A proper interpretation of higher order interactions can be better achieved in an empirical study guided by substantive theory rather than an exploratory study aiming to serve as a demonstration of a method. Also, it is a straightforward extension of the current model if the nonuniform DIF is of interest. To keep it clear and focused, our example only demonstrated the investigation of uniform DIF effect and their sources.



## Method

### *Data Source*

The sample for this study was comprised of 504,173 participants. Among them, 49.4% identified as boys and 50.6% identified as girls. All of these participants were 15-year-old students who responded to the PISA 2009 student survey. They represented students at that age from 71 nations or jurisdictions. Some jurisdictions that participated in PISA 2009 got excluded from our analysis because we could not find their corresponding national-level developmental indices. The sample size for each nation or jurisdiction ranges from 329 to 38,250, with a mean of 7101.

### *Measurement*

**Reading Attitude Scale in PISA 2009 Assessment** The reading attitude scale examined in this study was part of the student questionnaire. The scale measured the enjoyment of reading, which reflects a positive attitude towards reading. This scale contained 11 items. Students responded to each item with a four-point response scale (i.e., ‘strongly disagree’, ‘disagree’, ‘agree’, and ‘strongly agree’). A simplified version of this scale is attached in Appendix 4.1 to demonstrate the content of each item.

**Human Development Index (HDI)** HDI is a composite index of life expectancy, education, and average income. It ranges from 0 to 1. A nation scores higher on HDI when its population has a longer life expectancy at birth, longer period of education, and higher average income. The HDI is viewed as an index of the potential that people can do the things they want to do in their life. For this study, the 2009 HDI index for each nation was obtained from the *Human Development Report 2010* (United Nations Development Programme, 2010).

**Gender Inequality Index (GII)** GII is an index measuring gender disparity. It is a composite measure of gender inequality in three areas: reproductive health, empowerment, and labor market participation. Values of the 2008 GII were also retrieved from the *Human Development Report 2010* (United Nations Development Programme, 2010). It ranges from 0, which indicates that women and men perform equally, to 1, which indicates that women have poorest opportunities in all measured dimensions. As the 2009 GII index was not reported in that report (United Nations Development Programme, 2010), the 2008 GII index was used in our study.



## Data Analysis

As described earlier, two-level logistic regression models were used to detect gender DIF and explore the possible sources of such DIF effects across nations. Student responses on the reading attitude scale range from 1 to 4, with 4 possible response categories. The level-1 models (person level) are ordinal logistic regression models for the analysis of DIF, which are similar to those proposed by Swaminathan and Rogers (1990) and Zumbo (1999, 2008). In the level-2 models (nation level), the regression coefficients (i.e., intercepts and slopes) from the level-1 models, which include the coefficient representing an items' DIF effect, are treated as random variables whose variation could be explained by certain contextual factors.

The two-level proportional odds ordinal logistic regression models were conducted using the program *HLM 6* (Raudenbush, Bryk, & Congdon, 2004). In our case, the level-1 dependent variable was the item rating (i.e., the survey response) for each of the items. The level-1 full model can be expressed by the following heuristic equation:

$$\text{Prob}(Y_{pq} = k) = \beta_{0q} + \beta_{1q} \times \text{grouping} + \beta_{2q} \times \text{ability},$$

where  $Y_{pq}$  is the probability of a respondent  $p$  choosing response category  $k$  on item  $q$ . Ability is used as the matching variable here. Total score or purified total score is commonly used as matching variable in DIF investigation. Grouping is usually a dummy variable used to indicate whether a person belongs to the group of interest (i.e., focal group) or belongs to the reference group. A statistically significant coefficient associated with the grouping variable in the model can be used to signal the presence of uniform DIF.

More formally, using HLM notation, the following equations were fit for each item. The notation  $i$  denotes level-one units (i.e., the student), and  $j$  denotes the level-2 units (i.e., the nations). See the *HLM 7 Manual* (Raudenbush, Bryk, Cheong, Congdon, & Toit, 2011, pp. 111–112) for descriptions of the other notations and estimation method.

The level-1 model is:

$$\begin{aligned} \text{Prob}[R_{ij} \leq 1 | \beta_j] &= \phi_{1ij}^* = \phi_{1ij} \\ \text{Prob}[R_{ij} \leq 2 | \beta_j] &= \phi_{2ij}^* = \phi_{2ij} + \phi_{1ij} \\ \text{Prob}[R_{ij} \leq 3 | \beta_j] &= \phi_{3ij}^* = \phi_{3ij} + \phi_{2ij} + \phi_{1ij} \\ \text{Prob}[R_{ij} \leq 4 | \beta_j] &= 1.0 \\ \phi_{1ij} &= \text{Prob}[\text{Rating}(1) = 1 | \beta_j] \\ \phi_{2ij} &= \text{Prob}[\text{Rating}(2) = 1 | \beta_j] \\ \phi_{3ij} &= \text{Prob}[\text{Rating}(3) = 1 | \beta_j] \\ \log[\phi_{1ij}^*/(1 - \phi_{1ij}^*)] &= \beta_{0j} + \beta_{1j}^* \text{grouping} + \beta_{2j}^* \text{ability} \\ &\vdots \\ \log[\phi_{3ij}^*/(1 - \phi_{3ij}^*)] &= \beta_{0j} + \beta_{1j}^* \text{grouping} + \beta_{2j}^* \text{ability} + \delta_0 \end{aligned}$$

In the level-2 models, the coefficients associated with the intercept and the slopes are formulated as random variables:

$$\beta_{0j} = \gamma_{00} + \mathbf{u}_{0j},$$

$$\beta_{1j} = \gamma_{10} + \mathbf{u}_{1j},$$

$$\beta_{2j} = \gamma_{20} + \mathbf{u}_{2j},$$

where  $\gamma_{x0}$  ( $x = 0, 1, 2$ ) are the average of the level-1 regression coefficients, and  $\mathbf{u}_{xj}$  ( $x = 0, 1, 2$ ) are random variables that represent the variations across nations. More specifically,  $\gamma_{10}$  is the coefficient associated with the average uniform DIF effect across nations,  $\mathbf{u}_{1j}$  is the variability among nations for the uniform DIF coefficient. When  $\mathbf{u}_{1j}$  is significant, we can add in nation-level factors to explain the variation of uniform DIF among nations.

## Results and Conclusions

To detect gender DIF, a two-level logistic regression model was run for each item separately. The level-1 (i.e., person-level) model was an ordinal logistic regression model with adjusted mean as the matching variable and self-reported gender as the grouping variable. The level-1 intercept and regression coefficients were allowed to vary at level-2 (i.e., nation-level) models. The first step was to detect if an average gender DIF effect exists, and if this DIF effect varies across nations. For this purpose, no predictor was added in the level-2 models. Table 4.1 presents a summary of the results. As shown in Table 4.1, all items except Item 5 were flagged as showing DIF when the significance of hypothesis testing of the regression coefficient for gender is used as the criterion. Given the huge sample size used in this study, it is not surprising that even a small difference can be statistically significant.

To help us understand and interpret the results, the odds ratio and its confidence interval were also reported for each of the regression coefficients of gender. The odds are a way of representing probability. The odds ratio is a useful indicator of the strength of the relationship. The odds ratio is 1 when there is no relationship. In Table 4.1, the odds ratios are calculated as girls' odds of endorsing that item relative to boys' odds. Thus, odds ratios that are higher than 1 suggest that girls having a higher chance to endorse the item compared to boys with the same level of overall reading attitude. Odds ratios lower than 1 indicate that the boys are more likely to endorse the item when their overall reading attitude is the same as the girls. For example, the odds ratio of the gender predictor is 1.36 for Item 8. This suggests that, on average, girls are 1.36 times more likely, than not, to endorse Item 8 compared to boys with the same overall scores on this reading attitude scale. The odds ratios of gender for items flagged as DIF ranged from 1.09 to 1.96. Most of the items had

**Table 4.1** Multilevel DIF models without level-2 predictors

Item	Gender DIF		Random effect	
	odds ratio <sup>a</sup>		S.D.	Chi-square (df)
	[CI]	who is more likely to endorse	(variance component)	p-value
1	1.09 ***	Girls	0.15	$X^2(71) = 740.03$
	[1.05, 1.13]		(0.02)	$p < .001$
2	1.12 ***	Boys	0.14	$X^2(71) = 474.81$
	[1.09, 1.16]		(0.02)	$p < .001$
3	1.12 ***	Boys	0.13	$X^2(71) = 524.00$
	[1.09, 1.16]		(0.02)	$p < .001$
4	1.19 ***	Boys	0.17	$X^2(71) = 904.39$
	[1.15, 1.25]		(0.03)	$p < .001$
5	1.00	–	0.13	$X^2(71) = 456.14$
	[0.97, 1.04]		(0.02)	$p < .001$
6	1.09 **	Girls	0.10	$X^2(71) = 308.01$
	[1.07, 1.12]		(0.01)	$p < .001$
7	1.41 ***	Boys	0.19	$X^2(71) = 1353.38$
	[1.35, 1.47]		(0.04)	$p < .001$
8	1.36 ***	Girls	0.11	$X^2(71) = 410.35$
	[1.33, 1.40]		(0.01)	$p < .001$
9	1.19 ***	Boys	0.13	$X^2(71) = 562.45$
	[1.15, 1.22]		(0.02)	$p < .001$
10	1.18 ***	Boys	0.09	$X^2(71) = 301.05$
	[1.15, 1.20]		(0.01)	$p < .001$
11	1.96 ***	Boys	0.30	$X^2(71) = 2293.27$
	[1.82, 2.08]		(0.09)	$p < .001$

Note: <sup>a</sup>For odds ratio within the range of 0–1 (i.e., DIF effect favoring boys), the reciprocal of the odds ratio was reported for the ease of interpretation

CI stands for Confidence Interval;

\*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$

small to moderate DIF effect sizes. As suggested by the values of odds ratios, the item with largest gender DIF effect was Item 11, with boys more likely to endorse this item compared to girls with the same level of reading attitude.

When logistic regression models are used for DIF detection, several rules have been commonly used to determine if an item is showing DIF. It is usually recommended to use a rule which takes effect size of DIF into consideration, especially under the conditions where the sample size is very large. Following this common practice, it may be desirable to use a blended rule to flag DIF items (Zumbo, 2008) when multilevel logistic regression models are used in DIF investigation. In a blended rule for flagging DIF items, the effect size of the grouping variable (i.e., DIF effect), as signified by an odds ratio is used in addition to the statistical significance test of the regression coefficient of the same variable. The cut-offs for the odds ratio may be different depending on the context and purpose of the DIF study.

**Table 4.2** Summary of multilevel DIF analyses with 2009 HDI Index and 2008 GII Index

Item	Step 1: DIF detection	Step 2a: DIF explanation with 2009 HDI		Step 2b: DIF explanation with 2008 GII	
	Gender (Gamma <sub>10</sub> )	Gender (Gamma' <sub>10</sub> )	HDI (Gamma <sub>11</sub> )	Gender (Gamma'' <sub>10</sub> )	GI (Gamma <sub>11</sub> )
1	(+)***	(+)**	(-)**	N	N
2	(-)***	N	N	N	N
3	(-)***	N	N	(-)**	N
4	(-)***	(+)***	(-)***	(-)***	(+)***
5	N	N	N	N	N
6	(+)***	N	N	(+)**	N
7	(-)***	N	(-)**	(-)***	(+)***
8	(+)***	N	N	(+)***	(-)**
9	(-)***	N	(-)*	(-)***	N
10	(-)***	N	N	(-)***	N
11	(-)***	N	(-)**	(-)***	(+)*

Note: 'Gamma' denotes  $\gamma$  in the earlier description of the HLM model; N denotes for non-significant regression coefficient; (+) denotes for positive regression coefficient; (-) denotes for negative regression coefficient

\*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$

After all of the items were examined for uniform gender DIF effect, items with a significant grouping effect (i.e., DIF effect) were further analyzed by adding either HDI or GII to potentially explain the significant grouping effect across nations. Note that four nations or jurisdictions, Hong Kong-China, Liechtenstein, Montenegro, and Serbia, were reported on the 2009 HDI but not on the 2008 GII. This difference leads to a slightly different sample size for the following analysis when HDI or GII was entered in level-2 separately as explanatory factors for the gender DIF effect.

As shown in Table 4.2, the second column from the left-hand side presents the regression coefficients from the models where level-1 (person level) was an ordinal logistic regression model and the level-1 coefficients were allowed to vary across the units at level-2 (i.e., nation level). No predictor was included in level-2 models. This was the model we used to detect DIF. Positive signs signify positive regression coefficients, and negative signs indicate negative regression coefficients. Given the coding we used for gender groups (girls =1 and boys =0), a positive and statistically significant regression coefficient suggests that girls are more likely to endorse that item. Meanwhile, a negative and significant regression coefficient suggests that boys are more likely to endorse an item.

To seek sources of gender DIF effect, HDI and GII were added to the level-2 models separately to moderate the relationship between gender and responses. We did not add both HDI and GII to the level-2 models at the same time because we wanted to see if they could each explain some variability in the gender DIF effect across nations separately. Also, the correlation between these two national-level indices was high (Pearson correlation =  $-.836$ ). Adding variables that are highly

correlated with each other into the same regression model may result in a collinearity issue. Therefore, HDI and GII were entered in the multilevel models separately. The 3rd and 4th columns in the middle of Table 4.2 presented the regression coefficients of gender (i.e., average DIF effect) and explanatory variable HDI in the models where HDI was used in the level-2 models to moderate the relationship between gender and item responses at level-1. The last two columns in Table 4.2 were the regression coefficients of the grouping variable gender (i.e., average DIF effect) and explanatory variable GII in the models where GII was used as a level-2 predictor explaining the relationship between gender and item responses.

From Table 4.2, we can see that adding national level predictors changes the relationship between gender and item responses while overall reading attitude is controlled. After adding HDI, the significant gender DIF coefficients became non-significant for eight out of ten previously significant items (see 3rd column). This suggests that, for some of the items, the average gender DIF effect observed on these items can be explained by the variability of the HDI of different nations. For Item 4, the direction of average DIF effect changed from negative (i.e., favoring boys) to positive (i.e., favoring girls) after controlling for national HDI values. For five of the items, HDI appeared to be a significant predictor that negatively associated with the gender DIF effect (see 4th column). This negative association suggests that the higher a nation's HDI value is, the lower the gender DIF effect is found on that item in the country or jurisdiction.

When GII was added as a national level explanatory variable, the significant gender DIF effect of two of the items (i.e., Item 1 and Item 2) became non-significant (see the 5th column in Table 4.2). For three of the items (i.e., Item 4, Item 7, and Item 11) showing gender DIF favoring boys, GII was positively associated with the gender DIF effect, which suggests that larger gender DIF effects associated with higher gender inequality in a jurisdiction. For Item 8 which showed gender DIF favoring girls, the relationship between GII and gender DIF effect was negative. It indicates that larger gender DIF effects associate with lower gender inequality. As lower gender inequality (GII) means women are more likely to have equal opportunities as men do, it shows the better opportunities girls have relative to men in a jurisdiction the larger the magnitude of this gender DIF effect favoring girls is. Although we expected that gender inequality, which was represented by GII, would more likely be the source of gender-related DIF, the results suggest that HDI was a good, if not better, explanatory variable for the gender DIF effects in our study. One possibility is that the two indices, GII and HDI, are highly correlated. Based on the data used in this study, the correlation between GII and HDI is  $-0.836$ . It is also possible that differences in the years when HDI 2009, GII 2008, and PISA 2009 assessment took place may relate to the differential performance of these two indices. As there is a larger time difference between GII and item response data than the time difference between HDI and item responses, GII may be a less accurate national contextual factor compared to the HDI.

## General Discussion

The test takers or respondents, and their cognitive process of item responding and test taking all happen within a context. This context includes not only biological constraints and affordances but also the setting, environment, and culture in which people and their minds reside. Contextual factors can be viewed as the background where cognitive processes happen; they may also interact with the cognitive processes. The important role of contextual factors in understanding cognitive process have been recognized and emphasized by many researchers in different areas (e.g., Bronfenbrenner, 1979; Pintrich, Marx, & Boyle, 1993; Shalley & Gilson, 2004). In large-scale standardized educational testing, direct evidence of cognitive processes is usually not available, but item responses, test performance, and some background information regarding the test takers are usually recorded. This is partly because the primary interest of carrying out such assessments is to evaluate test takers' item responses and test performance, and make inferences about test takers' ability. A problem with analyses focused on item responses and test performance is their inability to identify various processes underlying a test taker's performance on an item or a test. Individual test takers may use different strategies to respond to the same item, and these different strategies may or may not lead to the same response. Item responses and test performance are the end products of test takers' cognitive process of item responding and test taking. Naturally, all of the factors that may affect the cognitive process of item responding can affect the item responses and test performance. Our goals are to introduce the ecological framework of item responding as a way of framing enabling conditions to explain item responding, and to demonstrate the usefulness of using such an ecological framework and multilevel modeling strategy to tackle the relationships between contextual variables and item responding. In doing so, we also hope to push the boundary out to an array of contextual variables wherein response processes evidence can be gathered or presented in validation practice.

Test takers bring their social and culture background and the ecological of their lives into a test setting. This complex context can be described through a combination of a variety of variables. The ecological model of item responding and test performance that we discussed in this chapter emerges from the contextual emphases of item responding and aims to describe the complex context of item responding in a systematic way. Five major aspects of the contexts, including properties of a test, characteristics of students, school factors, context outside of schools, and social context, are organized into an ecological framework. It is important to note that we are not suggesting that only these five layers of contextual factors exist, nor that these are the only ones worth investigating. There could be more layers, and the relationships among all the layers are not necessarily nested. For example, some of the layers may be overlapping or crossing with each other. This ecological framework is intended to be used as a conceptual framework guiding the investigation of relationships between contextual variables and item responses.

In the example, we presented a study to demonstrate the use of multilevel logistic regression models in DIF investigation. The model used to investigate DIF with

higher level explanatory variables can be adapted and extended into many other forms. For example, this model can be expanded to investigate nonuniform DIF effects by adding an interaction term of proficiency and group. More than two groups can be compared within this regression model framework by using dummy codes for multiple group memberships. Contextual factors from different layers of the ecological framework, and their interactions, can be added in the model as explanatory factors of a DIF effect. Large-scale international assessments usually have a large number of participating jurisdictions or nations, which will allow such analysis. However, the model becomes complex and the interpretation of the model becomes complicated very fast with an increasing number of variables and levels of the model. Also, as demonstrated in previous studies (e.g., Balluerka et al., 2010, 2014), the models can be adapted to model the differential functioning of a group of items instead of individual items. It is important to remind ourselves that, in contrast to many of the conventional DIF analysis in the literature, the gender DIF investigated in our example is an average DIF effect across nations or jurisdictions. In other words, when no average DIF effect is detected, it is still possible that, within some of the countries, gender DIF exists.

To sum up, the ultimate goal of exploring sources leading to item response or score variation is to contribute to the validation of inferences made from test scores. The example presented in this chapter demonstrated the potential of exploring explanatory factors for item responses and test performance with an ecological model as a framework. This framework can be used to guide the investigation of testing mechanisms or mediators and moderators of situational or contextual effects. We hope that, with this conceptual framework, the progress in explaining DIF and the understanding of item response process and test performance will accelerate.

## Appendix 4.1: Reading Attitude Scale

All of the following questions are responded using 4-point Likert-type scales, ranging from strongly disagree to strongly agree.

**How much do you agree or disagree with these statements about reading?**

1	Read only if have to
2	One of favourite hobbies
3	Like talking about books with others
4	Hard to finish books
5	Feel happy if receive a book as a present
6	A waste of time
7	Enjoy going to bookstores or libraries
8	Read only to get information needed
9	Cannot sit still and read for more than a few minutes
10	Like to express opinions about books
11	Like to exchange books with friends



Note: These items are not in their original format as presented to the participants of PISA 2009. The wording of these items are simplified. They are presented here to demonstrate the content of each item

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Balluerka, N., Gorostiaga, A., Gómez-Benito, J., & Hidalgo, M. D. (2010). Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema*, *22*, 1018–1025.
- Balluerka, N., Plewis, I., Gorostiaga, A., & Padilla, J. L. (2014). Examining sources of DIF in psychological and educational assessment using multilevel logistic regression. *Methodology*, *10*, 71–79.
- Bronfenbrenner, U. (1979). Contexts of child rearing: Problems and prospects. *American Psychologist*, *34*, 844–850.
- Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
- Organization for Economic Co-operation and Development (OECD). (2010). *PISA 2009 framework: Key competencies in reading, mathematics and science*. Paris, France: OECD Publishing.
- Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research*, *63*, 167–199.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & Toit, M. (2011). *Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *HLM 6 for Windows [Computer software]*. Skokie, IL: Scientific Software International, Inc..
- Shalley, C. E., & Gilson, L. L. (2004). What leaders need to know: A review of social and contextual factors that can foster or hinder creativity. *The Leadership Quarterly*, *15*, 33–53.
- Stone, J., & Zumbo, B. D. (2016). Validity as a pragmatist project: A global concern with local application. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice* (pp. 555–573). Newcastle, UK: Cambridge Scholars.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.
- Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*, 53–75.
- United Nations Development Programme (UNDP). (2010). *Human development report 2010: The real wealth of nations: Pathways to human development*. New York, NY: Palgrave MacMillan.
- Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*, 443–464.

- Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28, 369–386.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Research and Evaluation/Department of National Defense.
- Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223–233.
- Zumbo, B. D. (2008). *Statistical methods for investigating item bias in self-report measures, [The University of Florence Lectures on Differential Item Functioning]*. Florence, Italy: Università degli Studi di Firenze. URL: [http://faculty.educ.ubc.ca/zumbo/papers/Zumbo\\_Univ\\_Firenze.pdf](http://faculty.educ.ubc.ca/zumbo/papers/Zumbo_Univ_Firenze.pdf)
- 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, NC: Information Age Publishing.
- Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, 5, 1–23.
- Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, 12, 136–151.

## Chapter 5

# Putting Flesh on the Psychometric Bone: Making Sense of IRT Parameters in Non-cognitive Measures by Investigating the Social-Cognitive Aspects of the Items

Anita M. Hubley, Amery D. Wu, Yan Liu, and Bruno D. Zumbo

As Zumbo (2007) and Hubley and Zumbo (2013) note, discussions of ‘validity’ and ‘validation’ are framed and shaped by the measurement and psychometric models employed, be they classical test theory, item response theory, factor analysis, or axiomatic scaling theory. Thus, measurement models are not neutral in the validation process (i.e., they have their own underlying values and assumptions) and their consideration is necessary for a fulsome discussion of validity (Zumbo, 2007, p. 54). In focusing on item response theory (IRT), it is important to highlight the distinction between the platonic structure of the mathematical objects that we use in IRT to conveniently describe its use in item analysis and IRT as psychological theorizing about item responses. IRT is a statistical model of ‘item responding’ by its very description and name. As each item may be considered a test, the insights to be gained from IRT may inform the validity of inferences made from item and test scores.

Under the IRT model, typically up to three parameters may be estimated to describe people’s response patterns: (1) *a*-parameter, which is the slope of the tangent line at the point of inflection on the item characteristics curve (ICC) indicating an item’s ability to discriminate among respondents, (2) *b*-parameter, which is the threshold value of an item that a respondent’s amount of the latent variable must exceed to endorse the item, and (3) *c*-parameter, which is the lower asymptote of the IRT function indicating the probability of a respondent with very little of the latent variable endorsing an item. A less-discussed fourth parameter *d*, the upper asymptote (Barton & Lord, 1981; Loken & Rulison, 2010) has been used to acknowledge the possibility that a respondent with a very high amount of the latent variable may

---

A.M. Hubley (✉) • A.D. Wu • Y. Liu • 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](mailto:anita.hubley@ubc.ca)

not endorse an item (e.g., Linacre, 2004; Loken & Rulison, 2010; Osgood, McMorris, & Potenza, 2002; Reise & Waller, 2003; Rupp, 2003; Waller & Reise, 2009). We wish to emphasize IRT practice as psychological theorizing about item responses wherein the item parameters are more than just symbolic letters; that is, item parameters carry psychological information about the interaction between the test taker and the characteristics of the item or task and hence provide a window into response processes as a source of validity evidence. The item parameters are essential to this process as they form the kernels of an IRT characterization of item responding. We therefore need to know what these item parameters represent in a psychological sense.

Our primary aim in this chapter is to put flesh on the psychometric bone of IRT. We argue that IRT could be used as evidence of item response processes rather than just as a statistical model, as is commonly seen in its use in item analysis. In its essence, the item response function of IRT characterizes the interaction of persons and items. That is, it captures that interaction of the person's level of the latent variable (i.e., theta) and four or fewer item parameters. In describing the logistic function (i.e., s-curve) of IRT, researchers tend to use mathematical symbols in a way that ignores the fact that these coefficients could represent, or be, a lens through which we can view a psychological response process and, as such, may inform test validation. This view of test validation is reflective of Messick's (1989, 1995) sense of substantive validity, which focuses on evidence about the process of responding (i.e., how and why people respond), as central to validation. It is also consistent with Zumbo's (2009) explanation-focused view of validation, in which validity is a contextualized and Pragmatic explanation for obtained test scores (Stone & Zumbo, 2016).

Roskam (1985) advocated for the importance of understanding the substantive meanings behind IRT parameters for psychological measures. He conjectured that items in personality inventories showed lower discriminating power when formulated in more general and abstract (i.e., less concrete) terms. Zumbo, Pope, Watson, and Hubley (1997) examined this conjecture empirically with the extraversion and neuroticism scales of the Eysenck Personality Questionnaire and showed that it did not hold. Generally, they found small to moderate nonsignificant relationships between item discrimination (i.e.,  $a$ -parameter) and both word and item abstractness with two and three parameter logistic (i.e., 2PL and 3PL) models; exceptions were significant positive correlations of .52 and .43 between item discrimination and each of word and item abstractness, respectively, for extraversion in the 3PL model only. None of these results supports Roskam's conjecture. In addition, Zumbo et al. reported that word and item abstractness showed significant positive correlations with item difficulty (i.e.,  $b$ -parameter) for extraversion in the 2PL model. All other correlations were statistically nonsignificant. It is noteworthy that the statistically significant findings were not consistent across either the 2PL or 3PL models of extraversion and neuroticism.

Only a few other researchers have explored the relationships between IRT parameters and item characteristics. Zickar and Ury (2002) examined relationships between two parameters (discrimination, difficulty) and item subtlety (vs. obviousness),

word frequency, frequency of misunderstood words, and social desirability with Goldberg's Adjective Checklist, a personality measure. They hypothesized that discrimination would be negatively related to the frequency of misunderstood words and item subtlety and positively related to word frequency. While discrimination was significantly negatively related ( $r = -.32$ ) to item subtlety, no significant relationships were found with frequency of misunderstood words or word frequency. They also hypothesized that social desirability would be negatively related to difficulty but no significant relationships were found with either parameter. They did find that social desirability and subtlety interacted to influence difficulty (but not discrimination), however. That is, social desirability of trait adjectives was negatively related to difficulty when adjectives were less subtle. No other significant relationships were found between the two parameters and the other item characteristics.

Although untested in their study, Zumbo et al. (1997) suggested that a high likelihood of endorsing an item when a person possesses very little of the latent variable in a personality measure might reflect social desirability in some circumstances (i.e., a positive relationship between the lower asymptote for an item ( $c$ -parameter) and social desirability of an item). Rouse, Finger, and Butcher (1999) tested this suggestion using the PSY-5 scales of the Minnesota Multiphasic Personality Inventory-II (MMPI-II) and found a positive relationship between the lower asymptote for an item and social desirability of an item, but only for some of the scales (e.g.,  $r = .60$  for Psychoticism,  $.49$  for Aggressiveness,  $.30$  for Neuroticism); the rest of the correlations were positive but small in magnitude. They further suggested that the relationship between social desirability and the  $d$ -parameter, the little studied estimate of the upper asymptote, should be examined as it was possible that individuals with a high level of the latent variable might not endorse an item because of its undesirability.

Reise and Waller (2003) questioned the findings of Rouse et al. (1999) and suggested that it would make more sense that there is a positive relationship between social desirability and item difficulty, although they did not directly test this. Based on their results using 2PL and 3PL models with the 15 factor scales from the MMPI Adolescent version (MMPI-A), Reise and Waller further suggested that large lower asymptotes may be the result of items that discriminate at one end of the latent variable only and may be due to the use of extreme modifiers such as 'always' or 'never' in items.

There is still much to be learned about the substantive meanings behind IRT parameters for non-cognitive measures as well as the relationships between each of these parameters and both the characteristics of items and how items are perceived by respondents. The purpose of the present study was to examine the relationship of five different social-cognitive aspects of items (i.e., wording specificity, availability heuristic, emotional comfort, meaning clarity, and social desirability) to IRT parameters estimated from responses to the Geriatric Depression Scale (GDS; Yesavage et al., 1983). Our goal was to extend the findings of previous research that attempt to add psychological meat to the psychometric bone when interpreting IRT parameters with non-cognitive measures.

## Method

### *Participants*

Two samples were used in this study. IRT parameters were estimated using the responses to the GDS from a sample of 729 adults (316 men, 413 women) ages 16–94 years ( $M = 55.1$ ,  $s = 20.56$ ) with education levels ranging from 2 to 21 years ( $M = 12.7$ ,  $s = 2.52$ ). A separate community sample of 30 men and women ages 21–88 years ( $M = 42.2$ ,  $s = 21.91$ ; see Table 5.1 for more information) completed the GDS items and provided ratings of the social-cognitive aspects of each item.

### *Measures and Procedures*

**GDS** The GDS is a 30-item measure of depressive symptoms in older adults, although the measure has also been used with individuals across the adult age range (e.g., Brink & Niemeyer, 1992; Zalsman, Weizman, Carel, & Aizenberg, 2001). Responses are made using a dichotomous (i.e., yes/no) format. Responses that reflect a more depressed response are scored as ‘1’ whereas responses in the non-depressed direction are scored as ‘0’. Items are summed and total scores can range from 0 to 30 with higher scores indicating greater severity of depressive symptoms.

**Social-Cognitive Aspects of Items** A sample of 30 adults independently completed the GDS items so they could rate each item on the degree to which: (a) the item wording was general vs. specific (wording specificity), (b) their ability to properly respond to the item took a short vs. long time (availability heuristic), (c) they felt uncomfortable vs. comfortable responding to the item (emotional comfort), (d) the meaning of the item was vague vs. clear to them (meaning clarity), and (e) most

**Table 5.1** Summary of 30 raters’ personal characteristics

	Raters (N = 30)
Males:females	16:14
Age (years)	$M = 42.2$ ( $SD = 21.91$ )
Education (frequency):	
Grade 10–11	1
Grade 12–13	4
College/University Incomplete	6
College/University Complete	13
Post-Graduate (Masters/Ph.D.)	6
Average GDS total score	$M = 13.53$ ( $SD = 4.00$ ), range = 7–23

**Rating Instruction**

You have two tasks. In the first task, please answer the question for how you felt *over the past week* by checking off YES or NO immediately beneath it. It is important to answer the question so that you may better complete the second task.

In the second task, please rate the question by circling a number for each of the five rating descriptions below it. You will be asked to rate 30 questions in total. For each question, you will complete the same set of 5 rating descriptions.

Meaning of numbers in rating descriptions:

- 3 = very
- 2 = somewhat
- 1 = slightly / a little
- 0 = neutral

**Example**

1. Does your mood often go up and down?

YES       NO

- The wording of this question was **general** 3 2 1 0 1 2 3 **specific**.
- To answer properly, I had to think a **short** 3 2 1 0 1 2 3 **long** time.
- I felt **uncomfortable** 3 2 1 0 1 2 3 **comfortable** responding to this question.
- This question was **vague** 3 2 1 0 1 2 3 **clear** to me.
- Most people would think responding YES to this question is *socially unacceptable* 3 2 1 0 1 2 3 **acceptable**.

*Note: please respond to the last rating task, even if you answered NO.*

Note. The above ratings were scored on a -3 to +3 scale.

**Fig. 5.1** Instructions for rating the GDS items

people would think selecting “yes” to the item would be socially unacceptable vs. acceptable (social desirability). In each case, a 7-point response scale (ranging from -3 to +3) was used. The instructions used in this task are provided in Fig. 5.1.

## Analyses

### *Fitting IRT Models*

Before calibrating the items for their IRT parameters, we checked the data assumption that the GDS responses were unidimensional. Judging by the eigenvalue >1 rule, a principal components analysis (PCA) of the tetrachoric matrix using the



FACTOR 7.02 program (Lorenzo-Seva & Ferrando, 2006) showed the presence of six factors. Using the same program, a parallel analysis based on the mean of random eigenvalues of marginally bootstrapped samples (PA-MBS; Lattin, Carroll, & Green, 2003) showed the presence of four factors. Nonetheless, because the first eigenvalue (6.42) was 3.5 times higher than the second eigenvalue (1.85), this suggests an essential unidimensional structure of the GDS items.

Five IRT models, with up to four parameters, were fit to the GDS response data provided by the sample of 729 adults: (1) the 1-parameter logistic model with  $b$  only (*1PL-b*), (2) the 2-PL model with  $a$  and  $b$  (*2PL-a.b*), (3) the 3-PL model with  $a$ ,  $b$ , and  $c$  (*3PL-a.b.c*), (4) the 3-PL model with  $a$ ,  $b$ , and  $d$  (*3PL-a.b.d*), and (5) a 4-parameter logistic model with  $a$ ,  $b$ ,  $c$ , and  $d$  (*4PL-a.b.c.d*).

The five IRT models for the present study could be specified in the most general form of the *4PL* model as follows:

$$P(X_{ij} = 1 | \theta_i, a_j, b_j, c_j, d_j) = c_j + (d_j - c_j) \frac{e^{1.7a_j(\theta_i - b_j)}}{1 + e^{1.7a_j(\theta_i - b_j)}},$$

where  $i$  denotes persons,  $j$  denotes items,  $X_{ij}$  denotes person's responses,  $\theta_i$  is the latent trait,  $a_j$  is the slope parameter,  $b_j$  is the threshold parameter,  $c_j$  is the parameter for the lower asymptote, and  $d_j$  is the parameter for the upper asymptote. The *3PL-a.b.c* model was obtained by fixing the  $d_j$  parameter in the above equation to be 1; the *3PL-a.b.d* model was obtained by fixing the  $c_j$  parameter to be 0; the *2PL-a.b* model was obtained by fixing the  $c_j$  parameter to be 0 and  $d_j$  parameter to be 1; and the *1PL-b* model was obtained by fixing the  $a_j$ , and  $d_j$  parameters to be 1 and  $c_j$  parameter to be 0 (see Loken & Rulison, 2010).

The OpenBUGS version 3.2.2 (Lunn, Spiegelhalter, Thomas, & Best, 2009) was used to calibrate the item parameters. OpenBUGS is a computer software program for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) estimation methods, which is the open source variant of WinBUGS. The Bayesian method was used for parameter estimation because it has been shown that Bayesian methods are more capable for estimating complex and heavily parameterized models where the likelihood is non-normal (Loken & Rulison, 2010). Prior distributions were selected for the parameter estimation (e.g., Baker & Kim, 2004; Mislevy, 1986). Specifically, we followed Loken and Rulison's specifications for prior distributions. The prior for the  $b_j$  parameter was set to be a normal distribution ( $N(0, 2)$ ). A lognormal prior ( $Lognormal(0, 0.125)$ ) was used for the  $a_j$  parameter, and the Beta distribution was used for the  $c_j$  parameter ( $Beta(5, 17)$ ), and for the  $d_j$  parameter ( $Beta(17, 5)$ ). The means of the posterior distributions of the item parameter estimates were reported.

Table 5.2 reports the descriptives of the item parameter estimates and fit indices of the five IRT models. The results showed that the *1PL-b* model showed the poorest fit to the data, as seen by the highest deviance information criterion (DIC) in the last column of Table 5.2. The two models without a  $c$ -parameter (*2PL-a.b* and *3PL-a.b.d*) fit relatively better than the two models with a  $c$ -parameter (*3PL-a.b.c* and

**Table 5.2** Descriptives of item parameter estimates and fit indices comparing five IRT models fitted to the GDS response data

Model	Parameter	Min	Max	Mean	SD	DIC
<i>1PL-b</i>	<i>b</i>	−.06	3.96	1.83	.98	17,420
<i>2PL-a.b</i>						
	<i>a</i>	.39	2.29	.91	.40	17,050
	<i>b</i>	−.04	2.38	1.37	.59	
<i>3PL-a.b.c</i>						
	<i>a</i>	.67	2.51	1.18	.44	17,210
	<i>b</i>	.45	2.39	1.54	.51	
	<i>c</i>	.01	.20	.07	.05	
<i>3PL-a.b.d</i>						
	<i>a</i>	<b>.47</b>	<b>2.50</b>	<b>1.00</b>	<b>.39</b>	<b>17,060</b>
	<i>b</i>	<b>−.39</b>	<b>2.35</b>	<b>1.02</b>	<b>.68</b>	
	<i>d</i>	<b>.69</b>	<b>.87</b>	<b>.79</b>	<b>.05</b>	
<i>4PL-a.b.c.d</i>						
	<i>a</i>	.77	2.80	1.34	.43	17,200
	<i>b</i>	.09	2.39	1.24	.55	
	<i>c</i>	.01	.19	.08	.06	
	<i>d</i>	.69	.87	.79	.04	

Note. *DIC* deviance information criterion. Lower *DIC* values suggest relatively better fit among the alternative models. Model selected as best-fitting for this data is bolded

*4PL-a.b.c.d*). The finding that the models with a *c*-parameter did not fit as well can be understood by the relatively small estimates of the *c*-parameters for the *3PL-a.b.c* model ( $M = 0.07$ ,  $SD = 0.05$ ) and for the *4PL-a.b.c.d* model ( $M = 0.08$ ,  $SD = 0.06$ ). Also, notice that the two better-fitting models (without the *c*-parameters) fit almost equally well to the data ( $DIC = 17,050$  vs.  $17,060$ ). Given that these two models fit the data equally well, we decided to report the results of the *3PL-a.b.d* model (highlighted in bold face in Table 5.2) because of the non-ignorable deviation (0.21) from the maximum value of upper asymptote of 1 ( $M = 0.79$ ,  $SD = 0.05$ ). Moreover, the *3PL-a.b.d* model, in contrast to the *2PL-a.b* model, provided the opportunity to study the *d*-parameter in relation to the social-cognitive aspects of the items. Table 5.3 reports the item parameter estimates for each item of the *3PL-a.b.d* model. The estimated *a*-, *b*-, and *d*-parameters were then treated as fixed values to study their relationships with the social-cognitive aspects of the items, as assessed by the 30 raters.

In addition to the three item parameters, the positive or negative phasing of an item to reflect a depression symptom was also treated as a fixed characteristic to study its relationship with the five social-cognitive aspects of item responding. Twenty GDS items are phrased negatively such that a ‘yes’ response reflects a depressed indicator or symptom (coded as 1; e.g., Do you feel downhearted and blue?); the other ten items are phrased positively such that a ‘yes’ response reflects a non-depressed indicator (coded as 0; e.g., Do you feel happy most of the time?).

**Table 5.3** 3PL-*a.b.d* IRT parameters for each of the 30 GDS items

GDS items	<i>a</i>	<i>b</i>	<i>d</i>
1. Are you basically satisfied with your life?	1.15	1.38	0.82
2. Have you dropped many of your activities and interests?	0.72	1.20	0.79
3. Do you feel that your life is empty?	1.44	1.52	0.81
4. Do you often get bored?	0.85	1.05	0.79
5. Are you hopeful about the future?	0.83	1.85	0.80
6. Are you bothered by thoughts you can't get out of your head?	0.86	0.24	0.83
7. Are you in good spirits most of the time?	1.40	1.77	0.83
8. Are you afraid that something bad is going to happen to you?	1.00	1.22	0.69
9. Do you feel happy most of the time?	1.45	1.38	0.85
10. Do you often feel helpless?	1.29	1.23	0.84
11. Do you often get restless and fidgety?	0.72	0.39	0.79
12. Do you prefer to stay at home, rather than going out and doing new things?	0.47	0.38	0.73
13. Do you frequently worry about the future?	1.06	0.50	0.82
14. Do you feel you have more problems with memory than most?	0.56	1.10	0.69
15. Do you think it is wonderful to be alive now?	0.91	1.99	0.76
16. Do you often feel downhearted and blue?	2.50	1.16	0.87
17. Do you feel pretty worthless the way you are now?	1.34	1.77	0.79
18. Do you worry a lot about the past?	0.86	1.52	0.72
19. Do you find life very exciting?	1.00	0.24	0.71
20. Is it hard for you to get started on new projects?	0.90	0.18	0.76
21. Do you feel full of energy?	0.88	0.02	0.84
22. Do you feel that your situation is hopeless?	1.23	2.35	0.77
23. Do you think that most people are better off than you are?	0.91	1.98	0.76
24. Do you frequently get upset over little things?	1.06	0.57	0.76
25. Do you frequently feel like crying?	1.10	1.45	0.78
26. Do you have trouble concentrating?	1.16	0.22	0.82
27. Do you enjoy getting up in the morning?	0.56	0.84	0.80
28. Do you prefer to avoid social gatherings?	0.49	0.89	0.73
29. Is it easy for you to make decisions?	0.74	0.62	0.83
30. Is your mind as clear as it used to be?	0.69	-0.39	0.86
Mean	1.00	1.02	0.79
SD	0.39	0.68	0.05

### *Social-Cognitive Aspects of the Items*

The descriptives of the ratings of the social-cognitive aspects of the items over the 30 raters and across the 30 items are provided in Table 5.4. Descriptives, presented for each item, can be found in Appendix A.

**Table 5.4** Descriptives of social-cognitive aspects of the items across 30 GDS items and over 30 raters

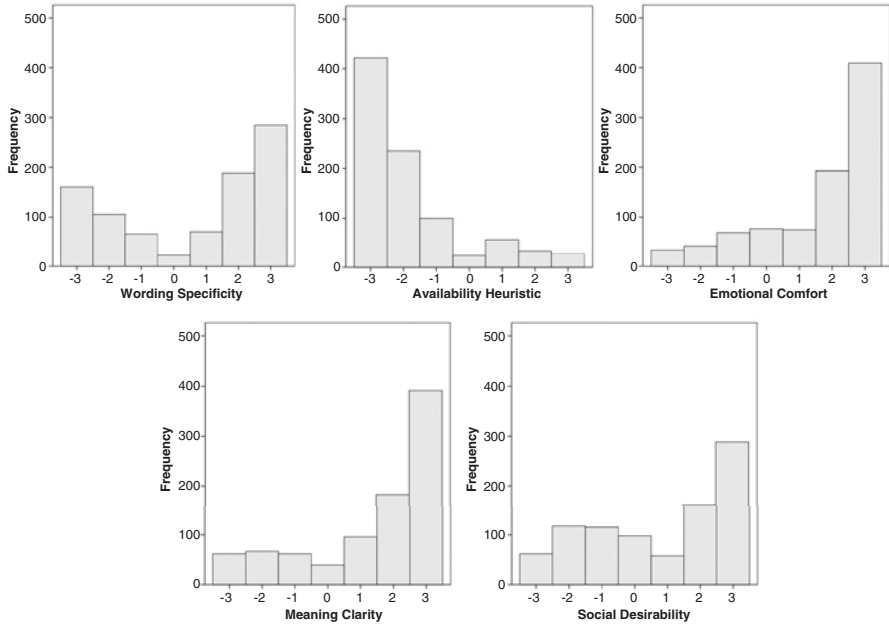
	Mean	Median	Mode	SD	Skewness
Wording specificity	0.60	2.00	3	2.359	-0.445
Availability heuristic	-1.81	-2.00	-3	1.614	1.519
Emotional comfort	1.59	2.00	3	1.762	-1.149
Meaning clarity	1.39	2.00	3	1.971	-1.042
Social desirability	0.79	1.50	3	2.077	-0.402

### *Associations Between IRT Parameters and Social-Cognitive Aspects of the Items*

The bivariate relationship of each parameter and phrasing of the items with the five social-cognitive aspects of the items was investigated using a simple regression analysis. Note that none of the five social-cognitive aspects of the items (rated on an ordinal scale of  $-3$  to  $3$ ) were normally distributed (see Fig. 5.2 and Table 5.4). As a result, the usual Gaussian distribution-based ordinary least squares (OLS) regression was not appropriate for such non-normal data. Furthermore, there was a nested structure in the data collection of the five social-cognitive aspects of the items. That is, each rater repeatedly rated the 30 GDS items, and thus the 30 ratings were dependent of one another within each rater (thus, the independence assumption of OLS regression was violated). We needed to account for the dependence when studying and testing the relationship between social-cognitive aspects of the items and item parameters. We also collapsed the seven rating categories into three categories (coded as  $-1$ ,  $0$ , and  $1$ ) to overcome the problem of zero cell counts resulting from a sample size of 30 raters and skewed ratings. We carefully grouped the response categories so that the order of the 3 categories was meaningful and interpretable.<sup>1</sup> The overall shape of the 3-category distributions was checked and remained unchanged from that of the original 7-categories shown in Fig. 5.2. To deal with the above concerns about the data, we used multilevel ordinal logistic regression to study the relationship between each parameter and each social-cognitive aspect of item responding.

A total of 20 two-level proportional odds ordinal logistic regression models were conducted using the program *HLM 7* (Raudenbush, Bryk, & Congdon, 2011). In each case, the level-1 dependent variable was the ratings for one of the five social-cognitive aspects of the items. For each of these dependent variables, four regressions were run separately with the three IRT item parameters ( $a$ ,  $b$ , and  $d$ ) and the phrasing of the item as the independent variable. In all cases, the level-2 models specify the random effects such that both the intercepts and slopes were treated as

<sup>1</sup>Using the anchors of wording specificity as an example, the original categories of somewhat general ( $-2$ ) and very general ( $-3$ ) were grouped and coded as  $-1$ . The original categories of neutral ( $0$ ), slightly general ( $-1$ ), and slightly specific ( $1$ ) were grouped and coded as  $0$ . The original categories of somewhat specific ( $2$ ) and very specific ( $3$ ) were grouped and coded as  $1$ .



**Fig. 5.2** Distributions of ratings of social-cognitive aspects of item responding over 30 items and raters

random (See Appendix B for more details). A statistically significant level-1 slope regression coefficient suggested a true simple relationship between a social-cognitive aspect of the items (e.g., wording specificity) and an IRT item parameter (e.g., *a*-parameter). Because there is no easily interpretable effect size measure for ordinal logistic regression, we also computed Spearman’s rho correlation as indication of the magnitude of the association.

## Results

Table 5.5 reports the results of the two-level proportional odds ordinal logistic regression for each social-cognitive aspect of the items.

### *Wording Specificity*

Wording specificity ratings were statistically significantly and negatively related to the *b*-parameter, difficulty. The more specific the wording of an item was, the less amount of the latent variable ‘depressive symptomatology’ was required to endorse an item. The effect size indicated that this was a small effect. Wording specificity

**Table 5.5** Results for 2-level proportional odds, random effects ordinal logistic regression

	<i>a</i> -parameter	<i>b</i> -parameter	<i>d</i> -parameter	Phrasing
Wording specificity				
$\beta$ -weight	-0.118	<b>-0.286</b>	-1.351	<b>0.448</b>
SE	0.212	<b>0.115</b>	1.224	<b>0.189</b>
<i>p</i> -value	0.581	<b>0.019</b>	0.279	<b>0.025</b>
Spearman's rho	-0.078	-0.127	-0.035	0.106
Availability heuristic				
$\beta$ -weight	<b>-0.680</b>	-0.189	-0.471	0.045
SE	<b>0.268</b>	0.157	1.622	0.199
<i>p</i> -value	<b>0.017</b>	0.238	0.774	0.821
Spearman's rho	-0.088	-0.041	0.009	-0.009
Emotional comfort				
$\beta$ -weight	-0.138	0.068	0.438	-0.271
SE	0.169	0.128	1.575	0.165
<i>p</i> -value	0.420	0.599	0.783	0.111
Spearman's rho	-0.068	-0.006	0.007	-0.069
Meaning clarity				
$\beta$ -weight	-0.060	-0.195	0.909	0.091
SE	0.233	0.143	1.602	0.132
<i>p</i> -value	0.800	0.184	0.575	0.495
Spearman's rho	-0.054	-0.088	0.013	0.021
Social desirability				
$\beta$ -weight	<b>-0.381</b>	<b>-0.283</b>	2.229	<b>-1.987</b>
SE	<b>0.138</b>	<b>0.116</b>	1.458	<b>0.300</b>
<i>p</i> -value	<b>0.010</b>	<b>0.021</b>	0.137	<b>&lt;0.001</b>
Spearman's rho	-0.109	-0.112	0.095	-0.408

Note. For interpretational ease, the original signs of all slope regression coefficients were reversed. Statistically significant results were highlighted in bold face. Spearman's rho correlation was reported as an indication of effect size

ratings were not significantly related to discrimination or the upper asymptote. Wording specificity ratings were statistically significantly and positively related to the phrasing of GDS items, with a small effect. GDS items that are worded to reflect a depressed symptom (e.g., "Have you dropped many of your activities and interests?") tended to be rated as more specifically worded than items that have a more positive stance (e.g., "Are you hopeful about the future?").

### *Availability Heuristic*

Availability heuristic ratings were statistically significantly and negatively related to the *a*-parameter, discrimination, with a relatively small effect size. The longer it took a respondent to answer an item, the less able the item was to discriminate

among respondents' levels of depressive symptomatology. Availability heuristic ratings were not significantly related to difficulty or the upper asymptote; nor were they related to the phrasing of GDS items.

### ***Emotional Comfort***

Emotional comfort ratings were not found to be statistically significantly related to discrimination, difficulty, upper asymptote, or the phrasing of GDS items.

### ***Meaning Clarity***

Meaning clarity ratings were also not statistically significantly related to discrimination, difficulty, upper asymptote, or the phrasing of GDS items.

### ***Social Desirability***

Social desirability ratings were statistically significantly and negatively related to both discrimination and difficulty, with small effect sizes. The more socially desirable raters thought it was to answer 'yes' to an item, the less able an item was to discriminate among respondents' levels of depressive symptomatology. In addition, the more socially desirable raters thought it was to answer 'yes' to an item, the less amount of the latent variable 'depressive symptomatology' was required to endorse an item. Social desirability ratings were not significantly related to the upper asymptote.

Social desirability ratings were statistically significantly and negatively related to the phrasing of the GDS item, with a moderate effect size. Specifically, items that are worded to reflect a more depressed symptom (e.g., "Do you feel that your life is empty?") tended to be rated as less socially desirable to which to respond 'yes' than items that have a more positive stance (e.g., "Are you in good spirits most of the time?").

## **Discussion**

The present findings contribute to the rather small literature that attempts to provide further psychological meaning to the interpretation of IRT parameters, particularly in the case of non-cognitive measures. Most of the social-cognitive aspects of the items included here (e.g., availability heuristic, emotional comfort, meaning clarity)



are new to the literature. Among the five social-cognitive aspects of the items investigated in this study, the ratings of emotional comfort in responding to an item and meaning clarity were found to be unrelated to any of the  $a$ ,  $b$ , or  $d$  parameters.

In this study, we consider two different types of processes. First, we have raters' evaluations of different aspects of the items; these ratings could provide us with some evidence about construct irrelevant variance (e.g., social desirability, item misinterpretation). Second, we also use these ratings to try to understand what aspects of an item play a role in the interaction of that item with a respondent, which subsequently results in a particular response. In this study, we found the  $a$ -parameter, or discrimination, to be negatively related to ratings of the availability heuristic and social desirability. Items that discriminated better among individuals along the latent variable of depressive symptomatology were those that were rated as quick to respond to and less socially acceptable to endorse. How quickly a respondent feels that he/she could select a response may reflect a variety of variables. Some of these variables may include how clearly an item is worded, how clear the meaning of the item is, and one's comfort in responding, although these seem not to be the cause given that we found no association between the  $a$ -parameter and wording specificity, emotional comfort, or meaning clarity. Wording specificity and meaning clarity were included here to better understand the concept of item abstractness that has been used in previous research and break it down into what we viewed as its two key components. The present study finding is in line with the general findings of nonsignificant correlations between discrimination and each of word and item abstractness for 2PL and 3PL models with measures of extraversion and neuroticism in Zumbo et al. (1997) and nonsignificant correlations between discrimination and frequency of misunderstood words and word frequency in Zickar and Ury (2002). It is noteworthy that Zumbo et al. did find significant positive relationships between discrimination obtained with a 3PL model and each of word and item abstractness ratings for a measure of extraversion, which suggests that results might vary depending on the model used and the construct being examined. Nonetheless, the current findings provide further evidence that Roskam's (1985) conjecture that more general and abstract items would show lower discriminating power is not supported. Further research is needed to better understand the relationship between ratings of the availability heuristic and discrimination. Perhaps there are other qualities of the item that contribute to this relationship (e.g., specificity or uniqueness of the item to the construct being measured).

The present finding of a significant negative correlation between discrimination and social desirability ratings is contrary to Zickar and Ury's (2002) finding of no significant relationship between these two variables with a general personality measure. However, ratings of social desirability differ (i.e., whether rating social desirability of personality traits or of responses to items) and so results may differ based on the types of social desirability ratings used or the constructs of interest (e.g., extraversion, depression, self-esteem, reading comprehension). Thus, it will be important to further examine the relationship of social desirability ratings to discrimination in future research.

The threshold parameter  $b$ , or difficulty, was found to be negatively associated with wording specificity and social desirability in the present study. Zumbo et al. (1997) generally found nonsignificant relationships between difficulty and each of word and item abstractness; the only exceptions were significant positive correlations found between difficulty and each of word and item abstractness for extraversion in the 2PL model. Ratings of wording specificity and meaning clarity are new to the present study but were meant to assist in understanding word/item abstractness. The sign of the correlations are different in the present study and in Zumbo et al. due to the direction of anchors in the ratings but both results suggest that it requires a relatively higher threshold to endorse items with a more vague and less clear meaning on some, but not all, measures and with  $b$ -parameters obtained in some, but not all, logistic models.

Zickar and Ury (2002) reported a near-zero correlation between social desirability and the  $b$ -parameter obtained with a general measure of personality whereas, in the present study, a small but significant negative correlation was found using the present depression measure. Zickar and Ury further reported that difficulty and social desirability were negatively related when items were less subtle (i.e., more obviously or transparently measured the construct of interest) and positively related when items were more subtle. It would be interesting in future research to further examine the role of item subtlety and how it might interact with various social-cognitive aspects of items.

In the present study, we found that an upper asymptote ( $d$ -parameter) was a more appropriate parameter than a lower asymptote ( $c$ -parameter) for the GDS response data as evidenced by the lack of fit of models including the  $c$ -parameter and the notable deviation of the mean  $d$ -parameter from the maximum value of 1.0. This indicates that there were cases in which respondents with a high level of the latent variable of depressive symptomatology did not endorse an item in the depressed direction. In fact, the  $d$ -parameters for the GDS items ranged from .69 to .87. Our finding that the  $d$ -parameter is useful for modeling this data echoes recent advocates of the need for considering an upper asymptote when modeling responses on non-cognitive measures, such as clinical and personality instruments (e.g., Reise & Waller, 2003; Waller & Reise, 2009). Unfortunately, in this study, the  $d$ -parameter values did not show any statistically significant correlation with any of the five social-cognitive aspects of items. Further work is needed to understand the psychological meaning behind this under-studied parameter.

Phrasing of the GDS items was statistically significantly and positively correlated with wording specificity and negatively associated with social desirability. Thus, items phrased negatively (i.e., reflecting a depressive symptom) tended to be worded more specifically and were less socially desirable.

What is most striking about the findings to date about the relationships between IRT parameters and both the characteristics of items and how items are perceived by respondents is that (a) relatively little research has been conducted in this area despite the increased use of IRT with non-cognitive measures, and (b) when variables have been used in more than one study, there is often a lack of consistency in the findings. The lack of consistency in findings across studies may be a result of

insufficient power, differences resulting from the presence of different best-fitting (e.g., 2-PL vs. 3-PL) models across measures/studies, as well as different and perhaps improved methods over the past 15+ years used to estimate parameters and detect small associations as reported in the literature. In this study, we sampled 30 raters and carefully chose the analytical approach to model the non-normal and dependently rated data used here. It is also possible that the differences in findings may be because the relationships among IRT parameters and both the social-cognitive aspects of items and phrasing of items are more specific to the construct of interest than previously considered.

The most notable limitation of the present study may be the use of a sample to provide ratings of the social-cognitive aspects of items that is different from the sample providing the GDS data used to obtain the IRT parameters. This approach has been used in previous research (e.g., Zickar & Ury, 2002; Zumbo et al., 1997) and the two samples presumably come from the same population of men and women of different ages in the general community; however, as the ratings are subjective, it is possible that those provided by the sample of 30 raters might differ significantly from what might have been provided by the larger sample that provided the GDS data.

There are many directions in which future research might proceed. It would be useful for previous studies to be replicated to determine if their findings are robust. This research can also incorporate many of the social-cognitive aspects of items included here and in other studies (e.g., subtlety). Roskam's (1985) conjecture led to research that focused on personality measures. Given the possibility that findings may be specific to the particular construct being measured, different non-cognitive (e.g., clinical, forensic, social) measures besides personality measures should be considered. It is also worthwhile to pursue this type of research to a greater extent with cognitive or achievement measures given the long history of IRT use with such measures. For research that uses ratings reflecting how items are perceived by respondents, it would be interesting to design a study in which all of the item ratings are obtained from the same sample that provides the data used to compute the IRT parameters. It is a tremendous step forward to conduct this type of 'in vitro' research in which we consider the role of other variables (e.g., item ratings) in understanding IRT parameters. At some point, we could take a step even further to incorporate more 'in vivo' information (e.g., contextual, situational, cultural factors) (Zumbo, 2015) in understanding IRT parameters and evaluating the presence of construct relevant and irrelevant variance.

One can place our ideas in the historical context of IRT. Contemporary uses of IRT are focused on statistical estimation theory (either a variant of maximum likelihood or Bayesian methods) and use of the statistical parameters (i.e., item parameters) in test assembly, equating/linking, and test scoring. This is certainly adequate but it ignores IRT's history in twentieth century psychology and mathematics; for example, a case has been made that predecessors of IRT include Binet and Simon's (1916) measurement of children's mental age and L. L. Thurstone's (1925) method of scaling of psychological and educational tests (Goldstein & Wood, 1989). Both of these lines of research share a focus on relating the measurement of

interest (e.g., the latent variable  $\theta$  in IRT and chronological age in Binet and Simon's work) and the probability of correctly responding to (or endorsing) an item. Both Binet and Simon's work and Thurstone's work suggested an s-shaped relation that had psychological meaning. Later work in statistics and psychometrics by Lord (e.g., Lord, 1968, 1980), Birnbaum (1968), Tucker (1946), Bock (e.g., Bock, 1983) and others focused attention on the thorny statistical problem of quantifying and mathematically characterizing the item response function (i.e., the s-shaped function) relating the unobserved latent ability to the probability of correctly responding (or endorsing) an item. As Goldstein and Wood (1989) noted, what has been going on since the earliest IRT efforts (e.g., Ferguson, 1942) is item response modelling and not item response theorizing. That is, IRT has been about description of items rather than an explanation of why an individual selects a particular response to an item or task. Our aim has been to return us to a focus on IRT as a psychological theory about item responding.

In closing, there should be more discussion about "models" and "modeling" in validity, and their varieties of uses, meanings, and intentions (Zumbo & MacMillan, 1999). Even a brief glance at the psychometric literature points to the fact that, in validity research, the issue is less about a lack of models for new kinds of test data but rather a lack of awareness in the applied world that extant models can, and should, be used to provide a psychological lens through which we can better understand the response process. In other words, the nature of psychometric modeling needs to change to provide richer psychological interpretations of item analysis and responding that move beyond a platonic interpretation of mathematical symbols.

Appendices

Appendix A: Descriptives of Item Ratings for the 30 GDS Items Across 30 Raters

Item		Wording specificity	Availability heuristic	Emotional comfort	Meaning clarity	Social desirability	Item	Wording specificity	Availability heuristic	Emotional comfort	Meaning clarity	Social desirability
1	Mean	-1.267	-1.867	1.833	1.367	1.4	16	-0.067	-1.6	1.833	0.867	1.8
	Median	-2	-2	3	2	2		0.5	-2	3	1	3
	Mode	-3	-3	3	3	3		-3	-3	3	3	3
	SD	2.273	1.332	1.663	2.025	2.01		2.434	1.958	1.663	2.193	1.769
	Skewness	1.09	1.052	-1.308	-1.287	-1.221		-0.071	1.46	-1.212	-0.596	-1.516
2	Mean	-0.467	-1	1.5	0.379	0.267	17	1.2	-2.133	1.7	1.467	0.267
	Median	-1	-2	2	0	0		2	-2	2	2	-1
	Mode	-3	-3	3	-1	-2		2	-3	3	3	-1
	SD	2.047	1.857	1.717	2.111	1.964		2.124	1.279	1.601	1.978	1.999
	Skewness	0.298	0.519	-0.854	-0.149	0.156		-0.998	2.383	-1.039	-1.25	0.272
3	Mean	-0.733	-2.367	1.533	0.833	-0.267	18	1.067	-2.067	1.5	1.833	-0.533
	Median	-1.5	-3	2.5	1	-1		2	-2	2	3	-1
	Mode	-3	-3	3	3	-3		3	-3	3	3	-3
	SD	2.348	0.809	1.978	2.102	2.212		2.067	1.172	1.757	1.621	2.08
	Skewness	0.615	1.211	-1.213	-0.528	0.198		-0.723	1.65	-1.083	-1.534	0.408

(continued)

Item		Wording specificity	Availability heuristic	Emotional comfort	Meaning clarity	Social desirability	Item	Wording specificity	Availability heuristic	Emotional comfort	Meaning clarity	Social desirability
4	Mean	0.533	-2.3	1.9	1.7	0.867	19	0.767	-2.433	1.6	1.467	0.3
	Median	2	-3	2.5	2	2		2	-3	2	2	0
	Mode	2	-3	3	3	2		3	-3	3	3	0
	SD	2.315	1.208	1.539	1.705	1.943		2.402	0.774	1.812	1.871	1.915
5	Skewness	-0.474	2.388	-1.463	-1.553	-0.374	20	-0.537	1.436	-1.44	-1.045	0.14
	Mean	0	-1.6	1.8	1.067	1.867		0.533	-2.067	2.033	1.1	2.233
	Median	0	-3	2	2	3		1.5	-2	2	2	3
	Mode	-3	-3	3	3	3		3	-3	3	3	3
6	SD	2.546	2.094	1.71	2.196	1.756	21	2.315	1.143	1.033	2.057	1.305
	Skewness	0	1.333	-1.662	-0.928	-1.665		-0.403	1.475	-0.673	-0.602	-1.862
	Mean	0.967	-1.333	0.867	1.1	-0.167		1.1	-2.333	1.828	1.967	2.133
	Median	2	-2	2	2	-0.5		2	-2.5	2	2	3
7	Mode	3	-3	2	3	-3	22	2	-3	3	3	3
	SD	2.205	1.9	1.889	2.006	2.245		2.187	0.884	1.49	1.474	1.502
	Skewness	-0.72	0.742	-0.255	-0.559	0.105		-0.942	2.02	-1.145	-2.087	-1.876
	Mean	0.533	-1.833	2.133	1.167	1.967		0.8	-1.6	1.367	1.233	-0.367
8	Median	1.5	-2.5	3	2	3	23	2	-3	2	2	-1
	Mode	3	-3	3	3	3		2	-3	3	3	-2
	SD	2.432	1.744	1.548	2.167	1.847		2.188	2.094	2.025	2.128	2.205
	Skewness	-0.46	1.776	-1.673	-0.905	-1.811		-0.743	1.309	-0.993	-0.948	0.502
	Mean	0.033	-1.033	0.633	0.567	0.167	23	0.6	-1.833	1.633	1.267	0.333
	Median	0	-2	1	1	0		2	-2.5	2	2	0
	Mode	3	-3	3	3	2		2	-3	3	3	3
	SD	2.684	2.251	2.236	2.445	1.802		2.372	1.783	1.81	2.212	1.936
Skewness	-0.004	0.763	-0.437	-0.321	-0.115		-0.565	1.844	-1.278	-1.018	0.164	

9	Mean	0.067	-1.667	1.2	0.967	1.733	24	1.067	-1.8	1.167	1.533	0.167
	Median	1	-2	2	2	3		2	-2	2	2.5	0
	Mode	3	-3	3	3	3		3	-3	3	3	-1
	SD	2.532	1.749	2.091	2.414	1.999		2.273	1.69	2.001	2.08	2.001
	Skewness	-0.053	1.273	-0.721	-0.762	-1.576		-0.843	1.869	-0.964	-1.245	0.059
10	Mean	0.633	-1.8	1.233	0.9	-0.733	25	2.033	-2.267	1.567	1.933	0
	Median	1.5	-2	2	2	-2		3	-3	2	3	-1
	Mode	2	-3	3	2	-2		3	-3	3	3	-2
	SD	2.189	1.424	2.063	2.107	2.067		1.732	1.552	1.977	1.701	2.349
	Skewness	-0.552	1.004	-1.043	-0.736	0.847		-1.933	2.494	-1.382	-1.512	0.137
11	Mean	0.633	-1.8	1.233	0.9	-0.733	26	0.767	-1.533	1.367	1.767	0.167
	Median	1.5	-2	2	2	-2		2	-2	2	3	0
	Mode	2	-3	3	2	-2		3	-3	3	3	-1
	SD	2.189	1.424	2.063	2.107	2.067		2.373	1.907	1.956	1.832	1.967
	Skewness	-0.552	1.004	-1.043	-0.736	0.847		-0.482	1.19	-0.881	-1.612	0.101
12	Mean	1.1	-1.933	1.267	1.533	0.767	27	1.433	-2.1	1.9	1.7	2.067
	Median	2	-2	2.5	2	0.5		2.5	-2.5	2	3	3
	Mode	3	-3	3	3	0		3	-3	3	3	3
	SD	2.249	1.311	2.132	1.889	1.633		2.223	1.185	1.322	1.841	1.388
	Skewness	-0.932	1.147	-0.831	-1.141	-0.004		-1.256	1.672	-1.147	-1.301	-1.37
13	Mean	1.2	-1.7	1.667	1.8	0.633	28	0.9	-1.967	1.967	2.033	0.267
	Median	2.5	-2	2	2.5	0.5		2	-3	2.5	3	-0.5
	Mode	3	-3	3	3	3		3	-3	3	3	3
	SD	2.369	1.803	1.688	1.584	1.991		2.383	1.377	1.326	1.45	2.258
	Skewness	-0.905	1.596	-1.276	-1.376	-0.295		-0.709	1.126	-1.077	-1.88	0.01

(continued)



Item	Wording specificity	Availability heuristic	Emotional comfort	Meaning clarity	Social desirability	Item	Wording specificity	Availability heuristic	Emotional comfort	Meaning clarity	Social desirability
14	Mean	-1.833	1.667	1.4	1.133	29	1	-1.033	1.567	1.933	1.9
	Median	-2	2	2	1.5		2	-1.5	2	2.5	2
	Mode	-3	3	3	3		3	-3	3	3	3
	SD	2.427	1.44	1.647	1.993		2.117	1.938	1.794	1.574	1.348
	Skewness	-0.219	1.543	-1.705	-1.186		-0.888	0.933	-1.44	-1.872	-1.707
15	Mean	0.867	1.9	1.667	0.767	30	0.167	-1.3	1.633	1.367	1.633
	Median	2	3	3	1		0	-2	2	2	2
	Mode	2	3	3	3		2	-3	3	3	3
	SD	2.285	1.202	1.561	1.882		2.394	2.07	1.712	1.938	1.847
	Skewness	-0.772	1.014	-0.988	-1.41		-0.098	1.006	-1.326	-0.986	-1.422

## Appendix B

Using HLM notation, the following equations, as an example, specified the social-cognitive aspect of ‘wording specificity’ as the dependent variable and the  $\alpha$ -parameter as the independent variable. The notation  $i$  denotes level-one units (i.e., the items), and  $j$  denotes the level-2 units (i.e., the raters). See *HLM 7 Manual* (Raudenbush, Bryk, Cheong, Congdon, & Toit, 2011, pp. 111–112) for the descriptions of the other notations and estimation method).

### Level-1 (Item) Model

#### Level-1 Model

$$\begin{aligned} \text{Prob}\left[R_{ij} \leq -1 \mid \beta_j\right] &= \phi_{-1ij}^* = \phi_{-1ij} \\ \text{Prob}\left[R_{ij} \leq 0 \mid \beta_j\right] &= \phi_{0ij}^* = \phi_{-1ij} + \phi_{0ij} \\ \text{Prob}\left[R_{ij} \leq 1 \mid \beta_j\right] &= 1.0 \\ \phi_{-1ij} &= \text{Prob}\left[\text{SPECIFIC}(-1) = 1 \mid \beta_j\right] \\ \phi_{0ij} &= \text{Prob}\left[\text{SPECIFIC}(0) = 1 \mid \beta_j\right] \\ \log\left[\phi_{-1ij}^* / (1 - \phi_{-1ij}^*)\right] &= \beta_{0j} + \beta_{1j}^* (a_{-abd_{ij}}) \\ \log\left[\phi_{0ij}^* / (1 - \phi_{0ij}^*)\right] &= \beta_{0j} + \beta_{1j}^* (a_{-abd_{ij}}) + \delta_0 \end{aligned}$$

#### Level-2 Model

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \delta_0 & \end{aligned}$$

## References

- Baker, E. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.
- Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item response model*. Princeton, NJ: Educational Testing Service.
- Binet, A., & Simon, T.H. (1916). *The development of intelligence in children (The Binet-Simon Scale)* (E. S. Kite, Trans.). Baltimore, MD: Williams & Wilkins Co.

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison–Wesley.
- Bock, R. D. (1983). The mental growth curve re-examined. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 205–219). New York, NY: Academic.
- Brink, T. L., & Niemeyer, L. (1992). Assessment of depression in college students: Geriatric depression scale versus Center for Epidemiological Studies Depression Scale. *Psychological Reports, 71*, 163–166. doi:10.2466/PRO.71.5.163-166.
- Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika, 7*, 19–29.
- Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology, 42*, 139–167.
- 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.
- Lattin, J., Carroll, D. J., & Green, P. E. (2003). *Analyzing multivariate data* (pp. 114–116). Belmont, CA: Duxbury Press.
- Linacre, J. M. (2004). Discrimination, guessing and carelessness: Estimating IRT parameters with Rasch. *Rasch Measurement Transactions, 18*, 959–960.
- Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology, 63*, 509–525.
- Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989–1020.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers, 38*, 88–91.
- Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine, 28*, 3049–3082.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan Publishing Co. Inc.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.
- Mislevy, R. J. (1986). Bayes model estimation in item response models. *Psychometrika, 51*, 177–195.
- Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology, 18*, 267–296.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & Toit, M. (2011). *Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2011). *HLM 7 for Windows* [Computer software]. Skokie, IL: Scientific Software International, Inc.
- Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*, 164–184.
- Roskam, E. E. (1985). Current issues in item response theory: Beyond psychometrics. In E. E. Roskam (Ed.), *Measurement and personality assessment* (pp. 3–19). Amsterdam, The Netherlands: Elsevier Science.
- Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72*, 282–307.
- Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing, 3*, 365–384.

- Stone, J., & Zumbo, B. D. (2016). Validity as a pragmatist project: A global concern with local application. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice* (pp. 555–573). Newcastle, UK: Cambridge Scholars Publishing.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433–451.
- Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, *11*, 1–13.
- Waller, N. G., & Reise, S. P. (2009). Measuring psychopathology with non-standard IRT models: Fitting the four parameter model to the MMPI. In S. Embretson & J. S. Roberts (Eds.), *New directions in psychological measurement with model-based approaches*. Washington, DC: American Psychological Association.
- Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1983). Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research*, *17*, 36–49.
- Zalsman, G., Weizman, A., Carel, C. A., & Aizenberg, D. (2001). Geriatric Depression Scale (GDS-15): A sensitive and convenient instrument for measuring depression in young anorexic patients. *Journal of Nervous and Mental Disease*, *189*, 338–339. doi:10.1097/00005053-200105000-00015.
- Zickar, M. J., & Ury, K. L. (2002). Developing an interpretation of item parameters for personality items: Content correlates of parameter estimates. *Educational and Psychological Measurement*, *62*, 19–31.
- Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Psychometrics* (Vol. 26, pp. 45–79). Amsterdam, The Netherlands: 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, NC: IAP – Information Age Publishing, Inc.
- Zumbo, B. D. (2015, November). *Consequences, side effects and the ecology of testing: Keys to considering assessment 'in vivo'*. Keynote address, annual meeting of the Association for Educational Assessment – Europe (AEA-Europe), Glasgow, Scotland. <https://youtu.be/OL6Lr2BzuSQ>
- Zumbo, B. D., & MacMillan, P. O. (1999). An overview and some observations on the psychometric models used in computer-adaptive language testing. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 216–228). Cambridge, UK: Cambridge University Press.
- Zumbo, B. D., Pope, G. A., Watson, J. E., & Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement*, *57*, 963–969.

# Chapter 6

## Some Observations on Response Processes Research and Its Future Theoretical and Methodological Directions

Mihaela Launeanu and Anita M. Hubley

The current state in response processes research – that is, the scarcity of empirical studies on response processes, the methodological paucity with respect to examining these processes, the lack of theoretical models, and the nearly exclusive cognitive focus – is an invitation for us to critically reflect upon what factors might have led to this situation, and what are the implications of these circumstances on contemporary testing and validation practices. On the heels of this critical reflection, we will then: (a) explore alternative epistemological and methodological horizons for response processes research, (b) propose the expansion of the scope of this type of research, and (c) discuss potential new roles for response processes research.

### A Critical Evaluation of Response Processes Research

Response processes research has not yet flourished and achieved its full potential because of a sui-generis combination of factors such as: (a) restrictive overarching epistemologies, (b) limiting theoretical and practical frameworks underlying psychometrics and validity, and (c) a certain general climate in social sciences research. Together, these factors have led to a ‘failure to thrive’ in response processes research. This situation has significant consequences for the status of response processes in

---

M. Launeanu (✉)  
MA Counselling Psychology Program, Trinity Western University,  
7600 Glover Rd, Langley, BC V2Y 1Y1, Canada  
e-mail: [mihaela.launeanu@twu.ca](mailto:mihaela.launeanu@twu.ca)

A.M. Hubley  
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](mailto:anita.hubley@ubc.ca)

the day to day reality of validation work, for validity itself, and for developing and validating the inferences from, measures.

### ***Restrictive Overarching Epistemologies***

Applied sciences, such as psychology or psychometrics, are deeply rooted in the epistemological soil that has nurtured them. As Toomela (2009) noted, modern psychology is shaped by a strange mix of two difficult to integrate epistemologies: (a) the dominant Cartesian-Humean cause-effect epistemology that emphasizes a strict, linear model of efficient causality (experimental research), and/or association (correlational research) accompanied by intensive quantification, and (b) the Aristotelian structural-systemic epistemology that focuses on understanding the structures that underlie behaviours, and on final (teleological, dynamic) causation.

Response processes, defined as the mechanisms that generate observed test score variation (Embretson, 2010; Messick, 1995), may be considered underlying structures and processes that, through their very nature, would be more compatible research-wise with the structural-systemic (Aristotelian) epistemology than with the dominant Cartesian-Humean epistemology that is intrinsically focused on outcomes and inter-individual differences. However, given the supremacy of Cartesian-Humean epistemology over the last century, validity evidence based on response processes has been overlooked and the focus has been on computing correlations among the outcomes of these mechanisms (i.e., convergent, discriminant, test-criterion evidence).

### ***Restrictive Local Epistemological Frameworks***

The ‘breeding soil’ of response processes research has been situated within the epistemological headquarters of psychometrics. Traditionally, psychometric models and validity theories have been significantly influenced by several preferred epistemologies, sometimes conflicting with each other, but nonetheless similarly influential: (a) realism (e.g., theoretical entities or constructs, although unobservable, are real), (b) operationalism (e.g., theoretical entities are defined by how they are measured), (c) falsificationism and empiricism (e.g., data fit model approach), and (d) positivism and post-positivism (e.g., true or valid knowledge is accessible via rigorous observation combined with the analytical apparatus of logic and mathematics). These epistemologies tend to emphasize essentialist, objective, empirical, static, socially disconnected, outcome oriented, deductive, confirmatory, and decontextualized ways of knowing and conducting research.

Rooted in this type of epistemological soil, response processes research has closely mirrored these trends and promoted an exclusively decontextualized, mechanistic, confirmatory, atheoretical, essentialist, abstract, and socially disconnected

stance in the empirical investigations conducted in this area. This has led to several important shortcomings in this research domain: (a) significant limitations in scope (e.g., the almost exclusive investigation of intra-individual and cognitive response processes), (b) disconnection from the contextual influences that shape any testing situation, (c) minimization of idiographic and qualitative subjective experiences, and (d) lack of fruitful theoretical grounding (e.g., frameworks).

### ***Inclément Climate***

In addition to a relatively inauspicious theoretical soil, some trends in the general climate of social sciences research, in general, and in psychometrics, in particular, have also posed some limitations with respect to response processes research. First, the ‘cognitive revolution’ has led to an over-focus on investigating cognitive response processes at the expense of other relevant processes. Although such work is undoubtedly very important, it does not represent the entire domain of response processes. Second, largely atheoretical research practices inspired and supported by dustbowl empiricism and other incentives (e.g., the need to publish quickly) have led to a severe under-development of theoretical models in the social sciences in general, and in the response processes domain in particular. Third, the notable scarcity of research and research frameworks focused on context, processes, and mechanisms has led to ignoring contextual social response processes and to over-focusing on investigating the content of observed responses (i.e., what test-takers say) rather than the mechanisms underlying these responses. Fourth, in academic research in particular, over-specialization and the pressure to publish finite and nicely packaged empirical findings has encouraged a piecemeal approach to research, devoid of any robust integrative frameworks and cross-disciplinary dialogue. Psychometric work has not always integrated systematically substantive knowledge from various scientific domains, which has led to some excessively thin and, sometimes mechanistic, measurement models. And finally, the exponential development in statistical technology and the overreliance on statistical models at the expense of field research have encouraged an exclusively nomothetic approach and group aggregate analysis as well as the propagation of very restrictive definitions of constructs to fit with the conservative assumptions of most statistical models.

### ***Implications for Response Processes Research***

This state of the art has important implications for the status of response processes research in validation practice:

- (a) Response processes research has been the ‘Cinderella’ of validity practice, often left behind or ignored. Despite scholarly articles advocating for the

essential role of response processes as substantive validity evidence (e.g., Borsboom, Mellenbergh, & Van Heerden, 2004; Embretson, 2010; Zumbo, 2009), these processes have played a rather marginal and ancillary role in contemporary validation practices, and have been largely treated as an expendable and rather costly accessory, a sort of a ‘black box’ between test items and test scores, and, occasionally, as an ad hoc strategy for contextualizing data obtained from testing.

- (b) Test takers’ response processes have been largely ignored during the critical stages of test development and testing. Therefore, important information conveyed by response processes does not explicitly and intentionally inform any of the critical stages of test development and validation, or of an assessment or measurement process.
- (c) The generative, interpretive, and meaning making response processes involved in responding to test items or tasks, although essential for a correct understanding and interpretation of test findings, and vital for correct test use, are largely forgotten, dismissed or, at best, briefly discussed at the end of the testing process. In the relentless pursuit of an ever illusory objectivity, test takers’ subjective experiences as response processes are substituted by ready-made data in the form of test scores interpreted through an a priori lens or a definition of abstract, decontextualized constructs.
- (d) By over-focusing on finding evidence for an a priori determined meaning of a construct, the mainstream approach has systematically overlooked the person of the test taker, and the interaction between that person and the test items or tasks within a testing situation, as well as any possibility of novel information stemming from the testing situation itself. The answers on a test are seen as direct reflections or products of the construct of interest and not necessarily as meaningful constructions of the person who answered them in interaction with the broader testing context, including socio-cultural and idiosyncratic norms, and personal history. Whereas the construct has taken center stage in validity theory, the test taker who *is* the generator or creator of the test scores is often seen as a passive and largely irrelevant medium through which the abstract construct materializes into numerical test scores that can be later analyzed by various statistical models.
- (e) The impact of language and cultural dimensions on the response processes associated with responding to test items or tasks have been largely ignored, briefly discussed at the end of the assessment process, or addressed as item *bias* under a differential item functioning (DIF) quantitative framework. However, what may appear as bias under statistical inferential ‘logic’ may, in fact, be a crucial aspect of a culturally situated response process cycle not to be dismissed but rather further explored.

Given these shortcomings, new ways of conducting response processes research are necessary and we will discuss some proposed changes in the next section. These changes will have to address a new paradigmatic rapprochement by finding new epistemological alliances that can support and nurture response processes research,



new methodological strategies, an expanded definition and scope of response processes research as well as re-imagining the roles that this research may play in validation and test development.

## **Alternative Epistemological and Theoretical Grounding**

### ***New Epistemological Alliances and Alternative Paradigms***

An important step towards ‘enriching’ the present epistemological soil of response processes research is to examine some more nurturing paradigmatic and theoretical alliances that would further anchor and legitimize this type of research in the bigger picture of contemporary social research and practice. An in-depth review of each of these perspectives is beyond the scope of this chapter and the reader is encouraged to consult the suggested references for a more thorough perusal of these approaches.

Although the Cartesian-Humean and Aristotelian epistemologies seem fundamentally incompatible on many levels, they may be, in fact, complementary for building an integrated, holistic research perspective. With respect to response processes research, given its focus on processes and underlying structures and mechanisms, the structural systemic epistemology espoused by the Aristotelian epistemology may inspire some important future scientific strides in this domain. This would mean shifting the epistemological ground in validation practices away from an exclusive focus on experimental and correlational research on outcome measures and towards exploring the dynamic structures and processes that generate and underlie these outcomes.

To do this, we do not need to start ‘from scratch’ or ‘reinvent the wheel’ but rather to intentionally draw and build on the research methods and findings that have been around in psychology for some time, such as Vygotsky’s (1994) cultural psychology, Piaget’s (1972) genetic epistemology, or Lamiell’s (1987) idiothetic psychology. These theoretical approaches align well with the aim of investigating processes and dynamic structures using field data, and integrating intensive qualitative information, theory, and quantitative explorations, such as those required for examining response processes. Moreover, Lamiell’s idiothetic psychology provides an example of bridging the qualitative and idiographic aspects of response processes research with the nomothetic features explored in standardized testing, and may thus contribute to addressing the pressing need for integrating idiographic, individualized frameworks in assessment practices (Meyer et al., 2001).

Exploring and adopting alternative paradigms of investigating response processes may be salutary. For instance, the pragmatic paradigm (Long, 2013; Maxcy, 2003; Tashakkori & Teddlie, 2010) underlies mixed methods approaches and embraces situational and contextual influences. This focus permits a generous blend of experimentation within socially situated contexts that may support well the examination of interpersonal and ecological response processes. Adoption of such a

pragmatic paradigm may promote a fresher and more exploratory impetus in response processes research as well as more socially meaningful practices in psychometrics and validity.

The more recent ‘versions’ of empiricism, such as constructivist empiricism (van Fraassen, 1980) and structural empiricism (Maturana, 1990; Varela, Thompson & Rosch, 1991), as well as contextual pragmatism (Merten, 2013) are well-anchored empirically, and, at the same time, have adopted a socially situated, contextualized view with respect to knowledge creation and ways of knowing (Bruner, 1990). These characteristics may support furthering the research on situated, interactive, and collaborative-generative response processes that go beyond exploring the fit between observed and theoretically expected response processes.

### ***Explanatory Models: Generalization and Causation***

Generalization and causation are considered critical epistemological processes that are addressed, in one way or another, in every epistemology or theoretical explanatory model. In the next paragraphs, we propose that there are complementary variants of generalization and causation that may be particularly useful for examining and explaining response processes.

**Analytic Generalization** In contrast with extensive generalization, the key to analytic generalization is the use of “theoretical concepts to enable a more general perspective on specific qualitative patterns” (Halkier, 2011, p. 787). For instance, “in case study research the aim is not to consider the case as a sample of a larger population of like-cases but to discover *patterns and processes within the case*, and to use analytic generalization to extract the lessons learned” (Erickson, 2012, p. 687). In response processes research, this would mean mindfully recruiting substantive theoretical knowledge that would bring some general coherence and meaning to the patterns identified in the test takers’ scores, even in the absence of a full data set. Response processes research seems to be uniquely suited for using analytic generalization given the small sample sizes typically recruited in this type of research, the primarily qualitative nature of the data, and the quest for identifying and explaining the underlying mechanisms that could ultimately lead to building theoretical models of response behaviour in testing.

**Process Causation** The investigation of response processes fits well with a variant of causality called *process causation* that is primarily geared towards the study of *causal process* (Erickson, 2012; Maxwell, 2012). This is different from the most frequent type of causation implemented in science, in the sense that it uses field-based methods to study specifically and intently the actual array of events and actions that led to specific outcomes in *local* settings (Erickson; Maxwell). In essence, investigating response processes means illuminating the processual sequences that underlie responding to test items or tasks. In-depth observations of these sequences would be geared towards spotting the anomalies in the structure of

the qualitative or quantitative data (i.e., response patterns) in order to provide insights into the causal processes of interest.

**Agentic or Final Causation** Guba and Lincoln (2005) pointed to the paradoxical situation wherein “humans, being anticipatory, can produce an effect in anticipation of its cause” (p. 142). The idea of final causation or teleology is not new but rather a few millennia old (i.e., Aristotle). In the previous century, Heidegger (1975) also wrote extensively about “*causa finalis*” as a distinct type of causation that takes pre-eminence in human dealings due to human’s capacity for intentionality and anticipation. Agentic or final causation would be particularly relevant for investigating and explaining the agentic and teleological response processes that may shape test scores during the testing process.

### *Theoretical Models*

A careful integration of empirical knowledge and theory is crucial for building sound theoretical models given that theory building can go astray if previous substantive knowledge is not properly integrated into the models; a lack of empirical grounding may lead to empty and indefensible theories (Hesse-Biber & Burke Johnson, 2015). In response processes research, this is a critical issue given the extensive lack of theoretical grounding and/or theoretical models or frameworks.

A theoretical model of response processes could be built by integrating qualitative and quantitative information, and following a sequential interplay of inductive (data driven) and deductive (theorizing) stages. For example, formulating empirically grounded theoretical hypotheses from the observed response patterns in the qualitative data could be the beginning of the model building process. In the next step, relevant substantive knowledge may be mobilized to consolidate or alter these preliminary hypotheses, and new theoretically based hypotheses could be generated and further tested with additional qualitative or quantitative data. Alternatively, substantive knowledge may be the impetus or the starting point for generating hypotheses that can be further clarified using qualitative or quantitative data. Either way, a purposeful integration of theoretical and empirical rationales is critical for the building, testing, and refining of a model. This type of model building represents more than simply checking if the observed response processes fit with the theoretically expected response processes (AERA, APA, & NCME, 2014) as the emphasis is on progressively building knowledge by a back and forth interplay between theoretical rationales and empirical data.

Working exclusively from a top-down perspective (i.e., a theoretical model applied to data) may be too conservative and might constrict the richness of the data in this relatively new research domain. Working exclusively in an inductive fashion might result in piecemeal research with little to no explanatory or descriptive models. Therefore, a dialogue and mutual shaping informed by both strategies would be ideal. The inductive emergent trend will be best suited for understanding contextual,

situated response processes whereas the deductive approach would work best to mobilize existent theories for making sense of data. Building situated response processes models requires working at the intersection of multiple polarities or tensions: deductive-inductive, confirmatory-exploratory, and explanatory-descriptive. Hence, coding schemes should be flexible enough to (a) allow for a continuous scaffolding of processes as the researcher advances through the data and (b) resist premature “closure” for the sake of formulating *the* best fit model. A model should be theoretically informed but not so rigidly theory driven that new theoretical insights that might emerge from the empirical investigations are ignored or pushed aside.

## Methodological Horizons

Although the quest for explanation and explanatory models are laudable in any scientific endeavor and even the scientific aim per se, in response processes research there has been a rather premature and exclusive impetus towards formulating neat and tight explanations, either deterministic or contextual, even before earning a good, solid grasp of the multifaceted reality of response processes associated with responding to test items or tasks. However, very few empirical investigations of response processes have been conducted, and this domain is far from being thoroughly researched and understood at a robust descriptive level.

Therefore, eliciting rich descriptions of the response processes involved in testing via complementary qualitative methodologies would be helpful in order to better understand the subject matter, not only abstractly but also as a situated and socially relevant phenomenon. For example, discourse analysis (Derrida, 1982; Foucault, 1991), which is focused on eliciting meaning making processes from narratives while taking into account how language is shaping and building understanding and meanings, may be particularly suited for investigating response processes that emerge from test takers’ narratives during qualitative interviews about how test takers relate, and respond, to test items or tasks.

## *Relaxing the Confirmatory Impetus*

The idea that there is a conceptual yardstick (i.e., construct) against which one compares and measures the empirical data in order to explore the “best fit” represents a statistical reflex that may do more harm than good at this stage in response processes research. Achieving the best fit model works effectively in inferential statistics, but it has the potential to stall progress in the area of response processes research because it forces researchers to operate selectively and conservatively in a field with very little empirical data and almost no models to test. In our opinion, it is far too early to look for confirmations before we even know what are we looking for, and what are we going to find. Tukey (1993) stated, with respect to the state of

art in psychometrics: “Exploration has been rather neglected; confirmation has been rather sanctified. Neither action is justifiable.” (p. 822). This remark is a good reminder to keep the balance, and to temporarily relax the stringent confirmatory and explanatory discourses, at least until we know well enough the object of this type of discourse.

That being said, it is certainly not too early, but rather long overdue, to start generating hypotheses grounded in rich empirical data and supported by theoretical rationales, to build local models, and to experiment with these local models while bringing in substantive knowledge to support the conceptualization process. At this still incipient stage, more effort and attention should be dedicated to the generation of hypotheses and data exploration rather than examining if the empirical data fit with the construct theory. Although it is often forgotten, “hypothesis generation is a crucial stage of research because good experiments test specific and informative hypotheses” (Freedman, 2010, p. 49). Working exploratorily from the data to theory and back would be more helpful than focusing on testing models in a top-down, confirmatory fashion.

### ***“Thinking Outside the Q Boxes”***

We are past the time when we can afford to pit qualitative and quantitative approaches (‘the Q boxes’; Pearce, 2015) against one another. Any responsible and credible scientific endeavour must be fully conversant in both languages as scientific inquiry is essentially bilingual. For instance, obtaining substantial knowledge about the response processes involved in testing requires a purposeful immersion in the data given that “no amount of statistical maneuvering can get very far without a deep understanding of how the data were generated” (Freedman, 2010, p. 23). Therefore, integrating complementary qualitative and quantitative frameworks in researching response processes seems to promise the most solid results.

It is also critical to recognize that, from a quantitative standpoint, recent psychometric developments in the area of latent class mixture modeling allow for a more flexible inclusion of context in statistical models and for modeling qualitative heterogeneity and diverse response patterns clustered as latent classes (e.g., Zumbo et al., 2015). One way in which response processes research will benefit is to quantitatively model and test hypotheses generated from qualitative data using such frameworks.

Mixed methods designs lend themselves well to integrating apparently contradictory empirical findings or even epistemological stances, such as those that are likely to be encountered in response processes research. Mixed methods designs “enable a progressive reconfiguration of substantive findings and interpretations in a pattern of increasing insight and sophistication” (Caracelli & Greene, 1997, p. 23), and, therefore, may support the development of empirically grounded knowledge in connection with theory. These designs can support an ongoing integration of

quantitative and qualitative findings and make room for methods that may enrich our empirical and situated knowledge regarding response processes.

### ***Principled Discovery***

Another approach that may be particularly useful in response processes research at this stage is ‘principled discovery’ and the context-confirmatory approach. Principled discovery consists of methods that go beyond an initial planned hypothesis test to allow for emergent elaborations or refinements of those hypotheses (Hesse-Biber & Burke Johnson, 2015). Principled discovery involves at least two basic steps. First, the researcher carries out some form of exploratory analyses that may result in a finding that goes beyond the initial a priori hypothesis. This discovery may point to an underlying mechanism (i.e., mediator). Given the potential for being misled by statistical chance findings (due to multiple exploratory analyses), the second general step of principled discovery requires the researcher to seek some form of independent (or quasi-independent) confirmation of the discovery, either quantitatively or qualitatively, together with mobilizing relevant theoretical principles that may be relevant for making sense of that discovery. Julnes’ (1995, as cited in Wong, Wing, Steiner, Wong, & Cook, 2012) context-confirmatory approach uses the findings of principled discovery to infer an underlying mechanism. Response processes research may use such an approach so that, after discovering patterns in the qualitative data of test responses, researchers may employ theoretical principles and further empirical explorations to infer and formulate an understanding of an underlying mechanism or process.

## **Expanding the Scope of Response Processes Research**

In this section, we will explore the possibility of expanding the investigation of response processes beyond: (a) a cognitive focus, (b) an intra-individual focus, and (c) a theoretically expected focus, in order to examine (i) conative and self-referential, (ii) contextualized and situated, and (iii) emergent response processes.

### ***Beyond Cognitive Response Processes***

The most researched response processes related to test items and tasks are cognitive processes, such as item comprehension, interpretation, retrieval, and expression of a response to test items (e.g., Schwarz, 1999; Tourangeau, Rips & Rasinski, 2004), information processing and problem solving strategies (e.g., Embretson, 2010), and the cognitive operations involved in dealing with cognitive complexity (e.g.,

Embretson & Gorin, 2001). In addition, existing guidelines about how to study response processes (e.g., AERA et al., 2014) have focused on cognitive processes and have recommended cognitive interviewing as the method of choice to tap into these processes (Wills, 2005). Although there is no doubt that cognitive processes, such as understanding and interpreting test items, are critical and likely to be involved to a certain degree in all tests, cognitive processes are not the only type of response process that underlies the production of test scores, and that are relevant to the validity of inferences made from test scores.

When responding to test items and tasks, test takers are not only cognitively engaged with these items but also emotionally and motivationally engaged (e.g., Leighton, 2015). The process of generating test responses is strongly connected with evaluative and self-evaluative processes (e.g., importance, valence, and meaningfulness), motivational processes (e.g., motivations around impression management), reflexivity and self-referential processes, and meaning making processes, to name but a few.

A specific category of response processes that is primarily involved in responding to items on self-report measures are the self-referential processes. Given that self-report measures form the majority of tests in some areas of the social sciences (e.g., psychology), the lack of research about these response processes is concerning. The way in which test takers access self-related information and link information provided on test items to the self is qualitatively different from how test takers process the cognitive aspects of items and different from the response processes involved in problem solving (Berkovich-Ohana & Glicksohn, 2014).

Far from being a mere reflection or a causal product of a construct, the result or outcome of any assessment process emerges from the space between the item and a response, not solely in the test item, and not solely in the score (Bazerman, 1995; Markus & Borsboom, 2013). This *generative space* between test item and test score – including the interaction among the test-taker, item, and response option within a given context – represents the focus of response processes research. Test takers' subjective experiences, intentionality, and meaning making processes have been largely ignored by the stimulus-response (S-R) paradigm and by the “best fit” approach in testing. What has been, for too long, considered to be the black box of testing and validity may well be a treasure chest where we could find the key to a more accurate, contextual, and meaningful interpretation of test scores.

Making sense and making meaning are powerful human motivational processes. Testing is an intrinsically meaningful activity; it has a purpose, uses language and language interpretations to make sense of its findings, and is situated at the confluence of multiple value systems. In this sense, each test item or task can be conceived of as an invitation to meaning making, a sign around which a web of meanings is created through a series of iterative processes that attempt to decide which answer or response makes sense when certain contexts are invoked. Although meaning making processes are critical for testing, they have been the least explored and the least understood (Markus & Borsboom, 2013).

Everyone who has ever given or taken a test probably remembers the following frequent unsolicited comment in response to more general test items: “it depends”.



This can be seen as one of the markers of a contextualized meaning making process. Each respondent makes sense of the items by situating these items in a personally relevant context (e.g., by relating items to one's own beliefs, worldview, emotions, personal experiences, or in comparison to others one knows). In addition to the semantic understanding of test items, there is a *pragmatic* function of responding to test items or tasks that is connected to the context of testing. Moreover, there is no meaning or meaning making without a context (Leontiev, 2014); the human mind is constitutively contextualized (Edwards & Potter, 1992). This means that the context and contextualized response processes are central for the validity of inferences made from test scores, and not simply threats to validity, as *The Standards* seem to imply (AERA et al., 2014).

Therefore, it is important to either provide a definition of the context or to allow participants to define their field of meaning, and to focus on a uniform interpretation of test scores, not on uniform wording and administration given that the same word may be interpreted differently by test takers. Some methods that would allow investigation of the emergence of meaning when responding to test items are: the micro-genetic method (i.e., studying the emergence or 'genesis' of meanings while responding to test items; e.g., Wagoner, 2009), semiotic scaffolding (i.e., a systematic exploration of the hierarchy of meanings generated by test takers regarding certain constructs or items; e.g., Hoffmeyer, 2013), or methods of experience sampling in real-life settings (Mehl & Connor, 2012).

### ***Beyond an Intra-individual Focus Towards Intersubjective and Interactionist Frameworks***

One of the consequences of the severe individualistic bias within contemporary methodology is the absence of a sustained and interdisciplinary discussion regarding appropriate methodologies for studying intersubjectivity. Even in research on intersubjectivity, the unit of analysis is often the individual (O'Donnell, Tharp & Wilson, 1993). Nonetheless, testing represents an interactive situation in which the test taker interacts not only with the test items or tasks but also with a specific testing situation (e.g., individual versus group administration; for specific purposes that may be high or low stakes) and with the test administrator (or with remote administration as seen with computers or mobile devices). A series of motivational and evaluative processes are set in motion in this interactive situation. For example, test takers' scores can change significantly after experimentally manipulating the valences of the testing situation (e.g., Holtgraves, 2004). What happens in this interactive situation has substantial impact on the test scores and, further, on the interpretations of those scores.

In order to examine and understand situational response processes, we can use Mischel and Shoda's (1995) social cognitive-affective theory which states that:



Understanding individual functioning requires identifying the psychological situations that engage a particular person's characteristic/representative personality processes and the distinctive cognitions and affects that are experienced in them. Then, an individual's functioning should become visible in the distinctive or unique ways the person's behavior is changing across situations, not just in its overall level or mean of functioning (p. 674).

Testing represents precisely such a situation. Thus, it becomes important to identify what kind of psychological responses are evoked by certain test situations, and what are the situational and situated prototypic behaviours and responses engaged by test takers in those specific situations. The DIAMONDS taxonomy and measures such as the Riverside Situational Q-Sort (RSQ-8) that propose major dimensions of psychologically meaningful situation characteristics (e.g., duty, adversity, sociality) may be of particular use in such an endeavor (e.g., Rauthmann et al., 2014).

Furthermore, we can regard situational response processes as "stable but discriminative patterns of behaviors across situations or as unique bundles of temporally stable prototypic behaviours contextualized in psychological situations" (Mischel & Shoda, 1995, p. 674). In social cognitive theory, individual differences in patterns of behaviour across situations reflect underlying person variables such as an "individual's construals of their experiences, expectations, goals, values and self-regulatory strategies" (p. 675). These *interactive construals* are all critical in shaping test scores and in conveying meaning to test ratings, and they fit very well with a contextual, pragmatic explanation framework (e.g., van Fraassen, 1980; Zumbo, 2009).

Epistemologically, research on situated or contextual response processes can be conceptualized as an "idiographic analysis of behavioural coherence" across testing situations (Mischel & Shoda, 1995, p. 675). Thus, investigating situated response processes is critical to understanding intra-individual variability and patterns over time, and it represents a necessary complement to the nomothetic, aggregate based interpretations that most tests yield just by the way they were constructed. At the same time, a situated, interactive approach could avoid the danger of solipsism that ignores any universal, shared meanings and any possibility of exploring nomothetic relationships or laws.

In order to expand the domain of interactive response processes in testing, we can conceive of testing as a conversation or dialogue (Markus & Borsboom, 2013; Westerman, 2003). Rather than viewing the test as a detached observation of a construct, Markus and Borsboom (2013) proposed that the test be conceived of as a conversation, with test questions and answers forming a *sui generis* dialogue. Thus, we begin to see that testing as observation leaves test users in a passive stance of observing and recording test ratings and test users as interpreting the test scores as reflections of constructs. Testing as conversation involves an interactive attitude and a shared understanding of the items asked and answered. Dialogical analysis can be used to analyze response processes, and even Think-Aloud Protocol (TAP) data, as human meaning-making; dialogism recognizes that this information is imbedded in a socio-historical and cultural context, that the respondent and the test developer may not interpret the meaning of items or tasks the same way, and that there are consequences to this for test scores (Linell, 2009). Testing as conversation and

dialogical analysis requires a courageous and a rather uncomfortable relinquishing of validity and validation conceived on a predetermined foundation in favor of an indeterminate and open ended process.

### ***Beyond Inter-subjectivity Towards Socio-cultural, Socio-political, and Linguistic Frameworks***

The response processes associated with testing occur not only at intra-individual or interpersonal levels, but also within, and in dialogue with, the larger sociocultural context and shared linguistics frameworks that define or shape the meaning of test items and, implicitly, of test scores. In addition to individual, idiosyncratic meanings, there are meaning structures that are located within complex societal relationships, and, thus, responding to test items can be seen as social, cultural, and political ‘negotiations’ of meanings and actions (Gergen, 2009; Shotter, 1993). These meanings shape test takers’ responses to items and tasks as much as they shape test developers’ definitions and theories of constructs, and, hence, the interpretations of test scores are socially negotiated at the intersection of multiple realities. Test score interpretations are never only about the individual who responded to the test items or tasks but about that person as part of the broader social fabric. For example, the self-esteem construct does not have only an individual content and meaning but it is actually significantly shaped by socio-cultural dynamics that influence test takers and test developers alike (e.g., the pursuit of high self-esteem as a goal; difficulties viewed in terms of low self-esteem; self-esteem seen as a mandatory requirement for a ‘good life’ or as a sort of ‘social vaccine’).

Test developers’ meaning of a construct is not always shared by test takers. Therefore, it is not the examination of the test or task itself that is the most important for providing validity evidence but rather *what test takers make of it* and *what raters make of what test takers made of it*. Thus, an important focus in future response processes research is to examine the collaborative and sometimes shared meaning making processes that take place during testing, and how they impact the test scores. Moreover, meaning is not only in the test taker’s head, and does not depend exclusively on test takers’ interpretations, but “it is partly determined by (a) the structures of the world itself, and (b) the linguistic conventions of the social community” (Markus & Borsboom, 2013, p. 507). If we assume that meaning is not entirely in the head, then when we develop validity arguments we have to take into consideration the socio-cultural context and the response processes that stem from this context as much as we are willing to examine intraindividual response processes.

Bakhtin (1990) has noted that we ‘rent’ our words from their prior users and uses, extending them to those with whom we currently interact. Thus, the meaning of test scores cannot be separated from the linguistic conventions and interactions in

which it acquires its significance and relevance. Moreover, testing means being fluent in two languages (words and numbers), and being able to gracefully translate between them. The overwhelming majority of response processes underlying testing are linguistic in nature and, thus, it makes sense to investigate these linguistic mechanisms as part of response processes research.

If we approach testing as conversation, then we become interested not only in the expected response processes but equally interested in the emergent, situated, and constructed response processes that are seldom fully predictable from the beginning and from construct theory alone. Nonetheless, these emergent response processes may significantly impact the interpretations made from test scores, and, hence, need to be investigated by response processes research.

## **Expanding the Role of Response Processes in Validation and Test Development**

In the previous section, we focused on expanding what may be looked at or included in response processes research. In this section, we turn our attention to the role or function that response processes research may play in validation and test development.

### ***Exploratory and Explanatory Roles***

So far, the guidelines pertaining to response processes research have focused exclusively on examining whether the observed response processes match the theoretically expected response processes (AERA et al., 2014). Whereas this confirmatory focus is undoubtedly very important, it is equally important to recognize that, in testing, there may be relevant response processes that are not predictable from the beginning solely on the basis of construct theory and meaning. In many cases, we do not really understand what test takers are doing or thinking as they respond to test items or tasks. Investigating response processes can provide new information about test takers' subjective experiences and meaning making processes as well as the intra- and inter-personal dynamic of response processes during testing, and thus play more of an exploratory and explanatory role. We do not yet understand how response processes may remain consistent or vary across time and with repeated testing occasions. Nonetheless, any of these as of yet unpredictable and undiscovered response processes shape the test scores and may be equally relevant to the inferences made from test scores.

## *Changing the Unit of Analysis*

An important consequence of response processes research is that the focus in validity shifts from abstract constructs to the respondent and the context. This may also change the unit of data analysis from interchangeable individuals aggregated across various samples to the person in context. The rich, ‘thick descriptions’ (Geertz, 1973) brought in via response processes may enrich the contemporary ‘thin’ psychometric and validity discourses, and promote the development of ‘thicker’ psychometric frameworks able to accommodate the richness of new data. Furthermore, one could develop local, flexible, and contextualized theories (e.g., a theory of response behaviour formulated for a certain test within a specific context; Borsboom et al., 2004).

## *Using Response Processes in Experimental Approaches to Validation*

While some researchers may pursue more qualitative approaches to studying response processes, Bornstein (2011) has promoted an experimental based approach to studying response processes. Specifically, researchers can first identify the core response processes that occur during testing, and then manipulate these processes under various experimental conditions in order to determine the relative importance and role of these response processes across various testing contexts.

## *Psychometric Modeling of Response Processes*

Although a somewhat daunting task, some response processes could potentially be mathematically formalized and included in designing and testing psychometric models of item responses. For example, Markus and Borsboom (2013) suggested that the psychometric modeling of item wording, interpretations, or meanings would be a great stride forward in psychometrics. Zumbo et al. (2015) cited some research studies from the language testing domain that have addressed specifically this issue. This use of response processes is consistent with Embretson’s (2010) comment that: “The success of the construct modeling approach, especially for construct representation research, will depend on the ability of researchers and test developers to develop quantitative indices that define the theoretical mechanisms that are involved in the tasks” (p. 195). This suggestion highlights the possibility that some response processes may play an important role in psychometric modelling or model based measurement, if the identified underlying mechanisms of test takers’ responses could be quantified and estimated via psychometric models.

Notwithstanding this aspiration, it is understandable that some of the response processes may never be, and should never be, amenable to a mathematical formalization. For example, it is hard to envision that intrinsically indeterminate response processes (e.g., meaning making processes) could be fully mathematically modeled. Therefore, a critical question is whether, and to what extent could or should, response processes be quantitatively modeled within a psychometric model and at what level (e.g., intraindividual, contextual, or sociocultural).

### ***Development and Application of Process-Based Micro-Methodologies***

Response processes may also be viewed as a sui generis way to develop, pilot, and implement process-based micro-methodologies for test development and validation. For example, the thin slice prediction method that may predict future disintegration/decompensation in clinical psychology studies (Nalini & Rosenthal, 1992) may serve as an example of how response processes may become not just the object, but also the method, of investigation. Specifically, uncovering the micro-level response processes that test takers mobilize when responding to test items or tasks may be a way of investigating the dynamic aspects of various phenomena or experiences, and may help to develop process based investigative methodologies to be used in validation studies.

### ***Using Response Processes with Other Sources of Validity Evidence***

Response processes may also be able to contribute to validation work based on other sources of validity evidence, namely that of internal structure and test consequences. Specifically, different response processes may help explain the presence of minor factors, different factor structures, or a lack of measurement invariance among different groups or subgroups. Response processes research may become a fertile ground for articulating comprehensive consequentialist frameworks emphasizing social participation and responsibility as well as reflexivity in testing practice. Response processes research that takes into account test takers' experiences, values, and the interactions with the larger socio-cultural context would be uniquely suited to address the personal and social impact of test scores, and the consequences of testing in an integrated manner.

## ***Using Response Processes in the Test Development Process***

Finally, response processes may also play a key role in the test development process. An active exploration and examination of response processes during the test development, piloting, and refinement phases can help response processes act as regulatory, feedback mechanisms to assist test developers in selecting items that make use of intended response processes and avoid unintended or undesirable response processes. In this way, evaluation of response processes may contribute to the development and selection of more refined and better calibrated test items or tasks.

## **Concluding Statements**

Response processes should play a central role in validation and test development work because elucidating how test takers answer test items is critical in order to make accurate inferences based on test scores. Moreover, response processes research may support the development of contextualized and dynamic validation frameworks that take into account the situational or ecological variables of testing, which significantly impact the interpretations of test scores. For instance, investigating the constructive meaning making processes that occur during responding to test items could provide essential information for contextualizing and nuancing current validation practices. This represents a clear shift away from the supremacy of an abstract construct seen as the authoritative and definitive yardstick against which test scores are interpreted, towards re-focusing on the person (i.e., test taker) within a situational model of validation.

In order to accomplish its central role in validation work, response processes research has to move beyond the exclusive focus on investigating the intra-individual cognitive response processes and towards including non-cognitive, interpersonal, sociocultural, and meaning making response processes. Exploring alternative epistemological and methodological horizons in this area of research can provide the necessary boost to research studies conducted in this area and can significantly contribute to developing and implementing better validation frameworks and practices.

## **References**

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bakhtin, M. (1990). *Art and answerability: Early philosophical essays*. Austin, TX: University of Texas Press.
- Bazerman, C. (1995). *Constructing experience*. Carbondale, IL: Southern Illinois University Press.

- Berkovich-Ohana, A., & Glicksohn, J. (2014). The consciousness state space (CSS) – A model for a unified self and consciousness. *Frontiers in Psychology, 5*, 1–19.
- Bornstein, R. F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. *Psychological Assessment, 23*, 535–544.
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.
- Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.
- Caracelli, V. J., & Greene, J. C. (1997). Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 15*, 195–207.
- Derrida, J. (1982). *Eyes of the university: The right to philosophy*. Stanford, CA: Stanford University Press.
- Edwards, D., & Potter, J. (1992). *Discursive psychology*. London, UK: SAGE.
- Embretson, S. E. (Ed.). (2010). *Measuring psychological constructs: Advances in model-based approaches*. Washington, DC: American Psychological Association Books.
- Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343–368.
- Erickson, F. (2012). Comments on causality in qualitative inquiry. *Qualitative Inquiry, 18*, 686–688.
- Foucault, M. (1991). Questions of methods. In G. Burchell & P. Miller (Eds.), *The Foucault effect: Studies in governmentality* (pp. 73–86). Chicago, IL: University of Chicago Press.
- Freedman, D. A. (2010). Statistical models and shoe leather. In D. A. Freedman (Ed.), *Statistical models and causal inference: A dialogue with the social sciences* (pp. 45–62). Cambridge, UK: Cambridge University Press.
- Geertz, C. (1973). *The interpretation of cultures*. New York, NY: Basic Books.
- Gergen, K. (2009). *Relational being: Beyond self and community*. New York, NY: Oxford University Press.
- Guba, E. G., & Lincoln, Y. S. (2005). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (3rd ed., pp. 191–215). Thousand Oaks, CA: SAGE.
- Halkier, B. (2011). Methodological practicalities in analytical generalization. *Qualitative Inquiry, 17*, 787–797.
- Heidegger, M. (1975). *Poetry, thought, language*. New York, NY: Harper & Collins Perennial Library.
- Hesse-Biber, S. N., & Burke Johnson, R. (2015). *The Oxford handbook of multimethods and mixed methods research inquiry*. Oxford, UK: Oxford Library of Psychology.
- Hoffmeyer, J. (2013). *Why do we need a semiotic understanding of life? Beyond mechanism: Putting life back into biology*. Plymouth, UK: Lexington Books.
- Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30*, 161–172. doi:10.1177/0146167203259930.
- Lamiell, J. T. (1987). *The psychology of personality: An epistemological inquiry*. New York, NY: Columbia University Press.
- Leighton, J. P. (2015). *Accounting for affective states in response processing data: Impact for validation*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL, USA.
- Leontiev, D. (2014). Extending the contexts of existence: Benefits of meaning-guided living. In A. Batthyany (Ed.), *Meaning in existential and positive psychology* (pp. 97–114). Dordrecht, The Netherlands: Springer.
- Linell, P. (2009). *Rethinking language, mind, and world dialogically*. Charlotte, NC: Information Age Publishing.
- Long, D. M. (2013). Pragmatism, realism, and psychology: Understanding theory selection criteria. *Journal of Contextual Behavioral Science, 2*, 61–67.
- Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.



- Maturana, H. (1990). *Biology of cognition and epistemology*. Temuco, Chile: Ed Universidad de la Frontera.
- Maxcy, S. J. (2003). Pragmatic threads in mixed methods research for multiple modes: The search for multiple modes of inquiry and the end of the philosophy of formalism. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 51–89). Thousand Oaks, CA: SAGE.
- Maxwell, J. A. (2012). The importance of qualitative research for causal explanation in education. *Qualitative Inquiry*, 18, 655–661. doi:[10.1080/14733140112331385100](https://doi.org/10.1080/14733140112331385100).
- Mehl, M. R., & Connor, T. S. (2012). *Handbook for research methods for studying daily life*. New York, NY: Guildford Press.
- Merten, D. M. (2013). Mixed methods and the politics of human research: The transformative-emancipatory perspective. In A. Tasakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 135–164). Thousand Oaks, CA: SAGE.
- 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, 747–749.
- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, 56, 128–165.
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246–268. doi:[10.1037/0033-295X.102.2.246](https://doi.org/10.1037/0033-295X.102.2.246).
- Nalini, A., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences. *Psychological Bulletin*, 111, 256–274.
- O'Donnell, C. R., Tharp, R. G., & Wilson, K. (1993). Activity settings as the unit of analysis: A theoretical basis for community intervention and development. *American Journal of Community Psychology*, 21, 501–520. doi:[10.1007/BF00942157](https://doi.org/10.1007/BF00942157).
- Pearce, L. D. (2015). Thinking outside the Q boxes: Further motivating a mixed research perspective. In S. N. Hesse-Biber & R. Burke Johnson (Eds.), *The Oxford handbook of mixed and multimethod research*. New York, NY: Oxford University Press.
- Piaget, J. (1972). *The principles of genetic epistemology*. New York, NY: Basic Books.
- Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., & Funder, D. C. (2014). The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, 107, 677–718. doi:[10.1037/a0037250](https://doi.org/10.1037/a0037250).
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93–105.
- Shotter, J. (1993). *Conversational realities: Constructing life through language*. Thousand Oaks, CA: SAGE.
- Tashakkori, A., & Teddlie, C. (Eds.). (2010). *Handbook of mixed methods in social and behavioral research* (2nd ed.). Thousand Oaks, CA: SAGE.
- Toomela, A. (2009). How methodology became a toolbox – And how it escapes from that box. In J. Valsiner, P. Molenaar, M. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 45–66). New York, NY: Springer.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2004). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.
- Tukey, J. W. (1993). *Issues relevant to an honest account of data-based inference, partially in the light of Laurie Davies' paper*. Princeton, NJ: Princeton University.
- van Fraassen, B. C. (1980). *The scientific image*. Oxford, UK: Oxford University Press.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
- Vygotsky, L. S. (1994). The problem of the environment. In R. van der Veer & J. Valsiner (Eds.), *The Vygotsky reader* (pp. 338–354). Oxford, UK: Blackwell.



- Wagoner, B. (2009). The experimental methodology of constructive microgenesis. In J. Valsiner, P. Molenaar, N. Chaudhary, & M. Lyra (Eds.), *Handbook of dynamic process methodology in the social and developmental sciences* (pp. 99–121). New York, NY: Springer.
- Westerman, M. A. (2003). Quantitative research as an interpretive enterprise: The mostly unacknowledged role of interpretation in research efforts and suggestions for explicitly interpretive quantitative investigations. *New Ideas in Psychology, 24*, 189–211.
- Wills, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: SAGE.
- Wong, V. C., Wing, C., Steiner, P. M., Wong, M., & Cook, T. D. (2012). Research designs for program evaluation. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology, Research methods in psychology* (Vol. 2, 2nd ed., pp. 316–341). Hoboken, NJ: Wiley.
- 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, NC: Information Age Publishing.
- Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*, 136–151. doi:[10.1080/15434303.2014.972559](https://doi.org/10.1080/15434303.2014.972559).

# Chapter 7

## A Model Building Approach to Examining Response Processes as a Source of Validity Evidence for Self-Report Items and Measures

Mihaela Launeanu and Anita M. Hubley

One difficulty for many individuals conducting validity studies using the five sources of validity evidence based on the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) is that, for many constructs, it is not clear, or certainly not clearly documented in the literature, what response processes an individual might be expected to use when responding to test items or tasks designed to measure the intended construct. The pervasive absence of adequate theoretical frameworks or templates for building response processes models makes it difficult, if not impossible, to identify a priori response patterns and processes in order to integrate this knowledge into a coherent, testable explanatory model of test score variation.

In response to these difficulties, in this chapter we will: (a) provide some important conceptual clarifications regarding response processes, (b) propose a model of theoretically expected response processes when responding to self-report items related to self, (c) use the model to examine the response processes that were evident when individuals responded to items from the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965), and (d) discuss the implications of using response processes when conducting validity studies of self-report measures related to the self.

---

M. Launeanu (✉)

MA Counselling Psychology Program, Trinity Western University,  
7600 Glover Rd, Langley, BC V2Y 1Y1, Canada  
e-mail: [mihaela.launeanu@twu.ca](mailto:mihaela.launeanu@twu.ca)

A.M. Hubley

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](mailto:anita.hubley@ubc.ca)

## Conceptual Clarifications

Given some prevalent confusion in the literature with respect to what constitutes response processes, it becomes important to start this chapter by conceptually differentiating among observed responses, inferred response processes, and a response processes model, respectively. This conceptual differentiation bears important consequences for understanding construct validity, and for how we conduct validity studies for self-report measures using response processes inquiry.

### *Observed Responses*

Observed responses during testing represent either quantitative data in the form of test scores or qualitative data in the shape of respondents' spontaneous or probed comments about (a) the test items, (b) their thinking process, or (c) their choice of responses. Depending on the level of analysis of the observed responses (e.g., individual, interpersonal, or contextual), one may identify coherent observed patterns in the data either at the individual level or at level of interactive or emergent observed responses. If one queries further about how a respondent decided on a certain item response (e.g., "How did you choose that answer?"), one may elicit and then observe various steps, strategies, ways of reasoning, or problem solving skills that the respondent employed in order to arrive at a certain response.

All of these observed responses contribute to elucidating the *content* or the *what* of test takers' responses to test items as well as the various strategies they pursue in choosing a final answer to a test item. These observed responses mirror the main aspects of the construct targeted by the measure of interest, and, thus, may provide support for experiential experts' test content evidence as a source of validity. In this sense, the content of the observed responses is construct-specific; for instance, test takers' responses to a self-esteem measure would display content such as: self-worth, self-liking, and competency, which correspond to how the construct of self-esteem is defined, rather than content such as worry, social desirability or sociability, which might suggest alternative or competing interpretations of the construct being measured.

### *Inferred Response Processes*

Response *processes* represent "the theoretical mechanisms that underlie item responses" (Embretson, 1983, p. 179). Borsboom (2005) and Zumbo (2009) suggested that response processes represent the explanatory mechanisms of test score variation. It is important to note that, in these definitions, a response process refers to something that is inferred from the observed patterns of test responses, is further

explained by relevant theory, and represents a dynamic, unfolding mechanism (i.e., it has a temporal structure that may be concretized in discrete stages or pathways).

This differentiation between observed responses and inferred response processes is critical given that one of the most prevalent confusions in the empirical research on response processes has been that the same term, ‘response processes’, has been used interchangeably to denote two fundamentally different layers of a phenomenon: (1) the observed responses (e.g., test scores, interview comments, or the observed patterns of variation in test scores), and (2) the inferred response processes that represent the mechanisms underlying the observed test responses. Most empirical investigations that have asked respondents what they were thinking about while answering test items (i.e., cognitive interviewing), reported these results as response processes evidence. Therefore, most of the currently available empirical studies on response processes have been confined to reporting the content or the patterns in test takers’ observed responses as response processes evidence instead of investigating the mechanisms underlying these observed responses.

If we work inductively and post-hoc (i.e., after the test is built and used), and we start with the observed responses and information collected (e.g., via cognitive interviewing using Think Aloud Protocols (TAP) and/or verbal probing (VP)), then we could infer the presence of the response processes and explain how they work by mobilizing relevant theoretical assumptions and models.

Alternatively, if we specified the mechanisms that we expected to underlie responding to test items before the test items are built (i.e., a priori), then we would consider the observed response patterns to be concretizations of these underlying mechanisms. Ideally, for strong validity evidence, the model of the theoretically expected response processes would be specified a priori (i.e., before the test is administered and, better yet, in the test development phase). In practice, very few tests even mention response processes as source of validity evidence and, to our knowledge, no one has used an a priori model of response processes in validation.

**Invariant Processes Versus Construct Specific Observed Responses** The underlying mechanisms or response processes are likely to be invariant across several constructs or, in other words, they are construct non-specific, although a differential involvement may be noticed across different constructs. For instance, we anticipate that self-referential processing (i.e., accessing and processing information pertaining to characteristics of one’s self) may represent an invariant mechanism mobilized when people respond to self-report items that target various constructs related to one’s self. However, the observed strategies or approaches that respondents take in using that mechanism and the response contents observed may differ depending on the construct targeted by the self-report measure.

Specifically, although we anticipate that respondents will engage in self-evaluative self-referential processing when they answer self-report self-esteem, body image, or extraversion items, the approaches they take in selecting an answer (e.g., comparisons to others or to one’s past self) and the specific content would vary depending on the construct targeted by the self-report items (e.g., self-worth for self-esteem measures, weight and shape characteristics for a body image measure,

and sociability for an extraversion self-report measure). Strategies and content would be construct specific (i.e., uniquely or at least predominantly associated with a certain construct) whereas the mechanisms or processes underlying the strategies and content would be construct non-specific or invariant. That is, mechanisms or response processes would represent general socio-psychological processes recruited by various constructs (e.g., self-referential processing involved in responding to self-report items pertaining to any number of self related constructs).

Similarly, memory represents a general, fundamental psychological process involved in responding to self-report items sampling diverse psychological constructs (e.g., self-esteem, depression, extraversion, traumatic events). However, how exactly memory processes are involved and what kind of memory is mobilized while answering specific test items (e.g., autobiographical versus declarative) may be different across different measures, and this differential involvement of different memory processes may be relevant to the interpretations made from test scores. Using a metaphor, we can visualize the more general, construct non-specific processes as the bedrocks through which many streams of water (i.e., construct-specific contents and strategies) may run.

### ***Response Processes Model***

Finally, one of the main purposes for identifying response patterns and processes is to be able to integrate this knowledge in a coherent, testable, explanatory model of test score variation. In this sense, a response processes model represents a theoretical model embedded within a larger epistemological view that predicts, explains, or describes response processes underlying test score variation by using existing theoretical frameworks and/or substantive models and knowledge from various disciplines. Usually, a model requires theoretical intra- and inter-disciplinary integration. A response processes model answers the question of *why* respondents endorsed certain items or response options during testing.

A response processes model represents an integrative, epistemologically informed, multilevel framework that cogently articulates connections among (a) the construct specific and construct invariant processes (i.e., theoretical model), and (b) the situational-interactive and sociocultural context of testing (i.e., situated model). A visual representation of such a model is presented in Fig. 7.1, which illustrates the embedded multiple layers of a response processes model for responding to self-report items.

Although investigating all layers of the model is important in conducting validation studies, this chapter will deliberately focus only on elaborating the theoretically expected response processes. As an example or ‘proof-of-concept’, this theoretical model will be applied in an empirical study using the Rosenberg Self-Esteem Scale. The results of this study will be discussed in light of this theoretical model.

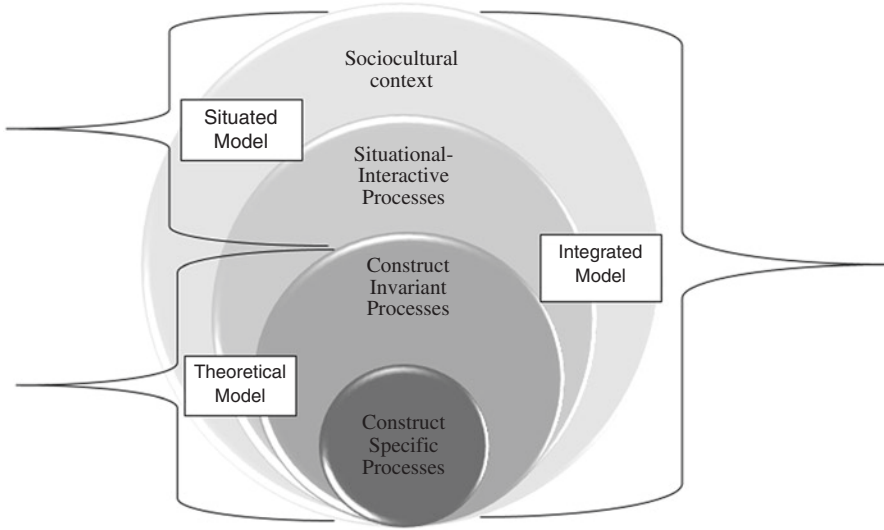


Fig. 7.1 Integrative multilevel response processes model

## Building a Response Processes Model for Answering Self-Report Items

In this section of the chapter, we will first discuss self-referential processing as a distinct neurological basis for processing information related to self (e.g., when responding to self-report items). Next, we will review self-evaluation and self-regulation as core response processes involved in responding to self-report items, and finally we will conclude by bringing this information together in a proposed model of response processes underlying self-report items.

### *Self-Referential Processing: The Neurological Basis of Self-Reporting*

Self-referential processing represents a distinct category of neuropsychological mechanisms that pertains to accessing, processing, and reporting information related to one's own person (Northoff et al., 2006). The term 'self-referential processing' is rooted in the neuro-imaging research tradition that has investigated brain networks and associated activity that support the human self and self related activities (Northoff et al., 2006). Therefore, self-referential processing can be seen as an overarching process that underlies the completion of any self-evaluative tasks such as responding to self-report items. This makes self-referential processing critical in understanding how people self-report information about themselves on surveys and questionnaires.

**Distinct Neural Signature** Magno and Allan (2007) found that there is “a common neural signature that is associated with self-referential processing regardless of whether subjects are retrieving general knowledge (noetic awareness) or re-experiencing past episodes (autonoetic awareness). Based on the neuro-imaging data, it seems that the neurophysiological correlate of self-referential activation is located in the medial prefrontal and parietal neocortical circuits” (p. 673). The existence of a neural signature of self-referential processing has been found in numerous research studies (Levine et al., 2004; Magno & Allan, 2007; Northoff & Bermpohl, 2004; Northoff et al., 2006), which suggests that neuropsychological mechanisms underlying processing and reporting of information about the self represent a category of response processes that is qualitatively distinct and thus are deserving of examination in validation studies using response processes inquiry.

**Sub-processes of Self-Referential Responding** Notwithstanding this neural signature, self-referential processes represent a complex system that involves a set of implicit and explicit subprocesses that may be cognitive (e.g., self-reflection) or affective (e.g., self-conscious emotionality such as shame or pride), as well as various sensory modalities and stimuli (e.g., images, auditory information) (Conway, 2005; Northoff & Bermpohl, 2004). Relatively distinct subregions within the neural network supporting self-referential processing correspond to specific subprocesses involved in self-referential processing (e.g., autobiographical memory, cognitive sub-processes, affective sub-processes), and recent research investigations discussed below have started to disentangle some of these sub-processes.

*Autobiographical Memory* Kim (2012) and Sajonz et al. (2010) distinguished two different neuro-pathways supporting self-referential processing and episodic memory retrieval, respectively. Furthermore, neuro-imaging and neuropsychological research studies with patients with neurological damage or neurodegenerative disorders have indicated that episodic and semantic (declarative) autobiographical memory retrieval elicited separate patterns of neural activity (Levine et al., 2004). In other words, it is possible that someone who has impaired episodic retrieval can still retrieve declarative or semantic self-knowledge in order to complete self-evaluation tasks. These research findings suggest not only that autobiographical retrieval may be the basis for self-referential processing, but also that autobiographical retrieval can rely both on declarative or semantic self-knowledge and on episodic memories of personal events, and that these two pathways are relatively independent from each other.

More recently, research studies have explored the role of episodic constructive or prospective retrieval as a form of autobiographical retrieval based on constructing possibilities and future scenarios using past episodic memories and present experiences (Kurczek et al., 2015; Kuzmanovic, Jefferson & Vogeley, 2016). Specifically, when responding to self-report items, test takers may imagine themselves in the future and decide on their final answer to test items based on that mental construction about future possibilities in which the self imagines itself to be involved. These self-referential processes have been denoted by terms like “prospective”, “constructive” or “anticipatory” response processes (Spreng, Mar & Kim, 2009), and they form an intrinsic part of processing information about the self when responding to self-report items.

*Cognitive-Semantic and Affective Sub-processes* Several research studies indicated that cognitive-semantic self-referential processing (e.g., self-reflection, cognitive self-appraisal) follows different neuropathways than affective processing (e.g., self-conscious emotionality, affective self-evaluation) (Mu & Han, 2010; Northoff & Bermpohl, 2004; Sui & Humphreys, 2013). This means that when individuals engage in responding to items related to self characteristics, they may employ differentially both cognitive or semantic processes and affective processes. Although neurologically and psychologically distinct, both categories of processes are organically intertwined whenever an individual processes information related to the self. It is noteworthy that these cognitive and affective pathways represent subprocesses of the overarching self-referential processing, and thus, they refer only to cognitive or affective processes that support processing information about self. Thus, cognitive self-referential processes, such as cognitive self-appraisal, are qualitatively different than task oriented appraisal or problem solving focused cognitive evaluations.

**Conclusion** Self-referential processing represents a distinct system of processing information about the self, supported by a set of cognitive-semantic and affective subprocesses. The basis of the self-referential processing is autobiographical retrieval either in the form of declarative semantic knowledge about the self (e.g., self-beliefs) or episodic (retrospective memory or prospective constructions) recall. Hence, self-referential processes warrant investigation of the role they play when a person responds to self-report items. This has important implications for an accurate interpretation of test scores from self-report measures; the degree of validity of the inferences made from these test scores may depend on the degree of differential activation of self-referential processes. This hypothesis is supported by the findings from an experimental research study on fake responding that indicated that fake, and thus invalid, responding only engaged cognitive and not self-referential response processes, which represent a neurologically different category of response processes (Holtgraves, 2004). Therefore, self-referential processing may play a distinct role in evaluating the substantive validity of interpretations based on self-report scores.

### ***Self-Evaluation and Self-Regulation***

Whereas self-referential processing denotes the overarching neuropsychological processes that are involved in any task that requires processing information about self-characteristics and that are captured via neuro-imaging (e.g., fMRIs, EEG, or ERP studies), self-evaluation and self-regulation represent core psychological response processes that can be inferred from observing how people respond to self-report items. Both self-evaluation and self-regulation are neurologically supported by the self-referential processing network of the brain but essentially they represent conscious psychological processes inferred from test takers' responses to self-report items. In responding to these items, individuals first evaluate themselves with respect to the self characteristic targeted by the self-report measure (e.g., self-esteem, body image, extraversion), and, depending on the result of this



self-evaluation (e.g., positive or negative self-esteem or body image), they may further engage in processes meant to regulate the potential discrepancy between the result of their self-evaluation process and their knowledge or expectation of self (i.e., self-regulation processes). This way, self-evaluation and self-regulation processes shape respondents' decisions regarding their final rating on self-report items.

**Self-Evaluation Processes** Self-evaluation response processes consist of several interconnected sub-processes that are involved in self-evaluative tasks: self-assessment, self-verification, self-enhancement, and self-improvement (Taylor, Neter & Wayment, 1995). All of these processes may be involved when responding to self-report items and may contribute differentially to the integrated evaluative judgment about one's self. The primary ways in which a person evaluates the self is via social comparisons (e.g., to others, based on feedback from others, based on one's social roles), self-standards comparisons (e.g., based on one's beliefs, expectations, and desires), and temporal comparisons (e.g., based on one's past or future performance and experiences). Self-evaluation processes mobilize both cognitive pathways (e.g., self-appraisals) and affective pathways (e.g., emotional evaluations accompanied by shame or pride) (Zell & Alicke, 2009).

Self-assessment, sometimes called self-appraisal in cognitive psychology, represents a predominantly cognitive sub-process meant to access information about the self as part of the self-evaluation process. An example of a self-assessment sub-process would be to remember the achievements that one has accomplished and then to decide, based on this information, that one is a competent person (i.e., the self-evaluative part). Self-assessment is often the first step in a self-evaluation process because it offers the basic information for other self-evaluative sub-processes, such as self-verification or self-improvement, to unfold.

Self-verification represents a very important self-evaluation sub-process that confers a sense of self-coherence and temporal continuity with respect to one's self. In other words, self-verification makes it possible for someone to recognize oneself as relatively the same person over time and across various situations. Maintaining a coherent sense of self across time and situations is critical to human beings' sense of self, and this explains why people are generally reluctant to change their self-ratings once they have decided on a rating and have a strong desire to appear consistent across items or over time. Self-verification can be inferred as the underlying process whenever individuals compare their present self with past or future versions of self (i.e., temporal self comparisons) or when they compare their self across various situations (i.e., situational self comparisons).

Self-enhancement represents a self-evaluative sub-process meant to ensure the maintenance of an overall favourable view of self. Preserving a positive self-image or high self-esteem represents a fundamental human need, and the process of self-enhancement selectively and purposefully allows individuals to focus on those self-characteristics that maintain this positive view of self while minimizing the relative importance of the self-characteristics that may threaten it (Somerville, Kelley, & Heatherton, 2010). For example, during self-evaluation, one may focus on one's excellent skills as a parent while minimizing the relative importance of the information about one's less developed housekeeping skills. In other words, people may

turn a blind eye to what they perceive to be their less desirable self-characteristics during self-evaluative processes in order to enhance their positive view of self.

Self-improvement as a self-evaluative process focuses on identifying possibilities to better one's self. Engaging in self-evaluation tends to predispose people to consider ways to improve their self-image or their perceived competency. Self-reflection about ways to improve one's self represents the building block of an important motivational process that keeps individuals engaged in the work of self-development and growth (Rahamim, Garbi, Shahar, & Meiran, 2016).

**Self-Regulation Processes** Self-regulation processes permit one to maintain a relative stability and integrity of self-identity as well as an acceptable homeostasis of self-esteem. That is, self-regulation processes maintain the consistency and coherence of self-representations and self-narratives through time (Hoyle & Wiley-Blackwell Online Books, 2013). The main categories of self-regulation processes that can be inferred from respondents' answers to self-report items are maintenance, restoration, compensation, and enhancing. Maintenance aims to preserve the already accomplished levels of self-esteem and self-integrity by dismissing negative feedback, removing self from a threatening situation, downplaying the impact of certain information, or by selectively focusing on positive information about the self. Restoration ensures the reparative process of injured self-esteem or self-integrity by engaging in various strategies such as: committing to self-improvement activities, engaging in value consistent behaviours, rationalization, and meaning making. Compensation facilitates engaging in actions meant to counterbalance the loss of self-esteem (e.g., working harder to compensate for a relational disappointment), whereas enhancing encourages individuals to explore strategies to increase one's self-worth and competence (Baumeister & Vohs, 2004).

Self-regulation and self-evaluative processes are tightly interconnected and inform each other. For example, if the result of self-evaluation represents a threat to one's self-esteem, then self-regulation processes meant to restore and protect an acceptable level of self-esteem are triggered.

Self-regulation processes also serve a significant motivational role as well in the sense that they stimulate the person to appropriate actions (Hoyle & Wiley-Blackwell Online Books, 2013). For example, when self-esteem is threatened, the individual may start ignoring negative feedback while asking only for positive feedback in order to protect and enhance self-esteem. Also, different sources of self-esteem may be accessed, and self-serving bias may also be mobilized as a self-regulation process meant to protect and enhance the self.

## *A Self-Report Response Processes Model*

Figure 7.2 depicts a model of response processes that are expected to underlie responses to self-report items.

Test items serve as stimuli that evoke or elicit autobiographical accessing and retrieval either in the form of declarative knowledge (e.g., general beliefs or attitudes

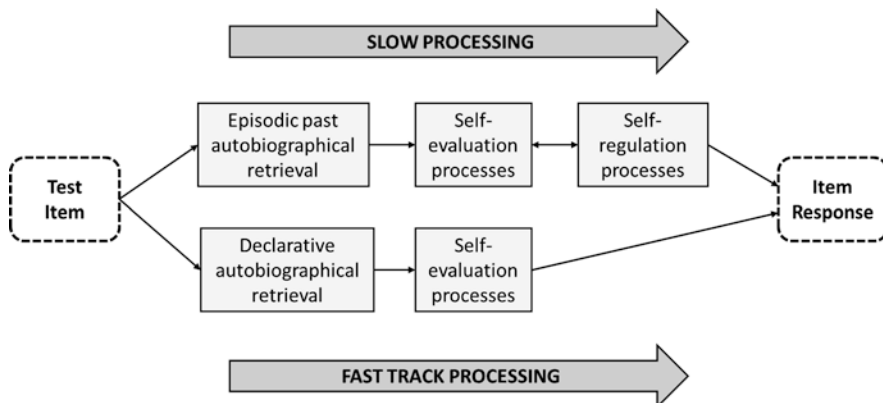


Fig. 7.2 Self-report response processes model

about/towards the self) or as episodic retrospective memory or prospective constructive processes. A series of self-evaluative and sometimes self-regulation processes are then set in motion that impact test takers' responses to items. Due to processing demands, episodic autobiographical retrieval is slower than declarative autobiographical retrieval. The item response is the outcome of these response processes supported by the neurological self-referential processing.

### Applying the Self-Report Response Processes Model to the Rosenberg Self-Esteem Scale

We next decided to apply the self-report response processes model in a study examining response processes as a source of validity evidence for inferences made from the widely known Rosenberg Self-Esteem Scale (RSES; Blascovich & Tamaka, 1993; Rosenberg, 1965). The RSES consists of 10 statements; half of the items are positively worded and half are negatively worded. The scale uses a four point agree/disagree Likert-type response format. First, we delineated the response processes that we expected to be evident theoretically when test takers respond to the RSES items. Then, we collected and analyzed data from a sample of 30 adults to examine these response processes. The findings of this study are discussed in light of the proposed theoretical model, and implications are outlined.

#### *Theoretically Expected Response Processes*

Theoretically, the RSES measures self-esteem broadly defined as a positive evaluation towards one's self with respect to a perceived sense of competence or mastery and a perceived sense of self-worthiness or personal value (Rosenberg, 1965;

Zeigler-Hill & Ebooks Corporation, 2013). Hence, it is expected that RSES scores and their descriptive qualities (i.e., content, direction, intensity, and stability) are the product of the underlying mechanisms of self-evaluation and self-regulation meant to maintain and enhance the ‘good self’. The results of these self-evaluation processes set in motion a series of self-regulatory processes (e.g., restore, protect, defend) which eventually determine the content, stability (stable or labile), direction (positive or negative), and intensity (degree from low to high) of the reported self-esteem, and, implicitly, the ratings for the RSES items.

Therefore, it was expected that, when responding to the RSES items, respondents will engage primarily in self-evaluation and self-regulation response processes pertaining to self-worth, competency, and self-liking. We also hypothesized that items that evoke self-esteem ‘critical moments’, such as rejection or failure, would trigger more intense self-evaluative and self-regulation processes. In engaging the basic self-process of self-evaluation, respondents would use social and temporal comparisons, as well as comparisons to personal standards, such as attending to their beliefs and expectations, to assess, verify, enhance, or improve their self-esteem (what others may refer to as critical self-evaluations). In making use of self-regulation processes, respondents were expected to selectively attend to feedback, use buffering, or access particular sources of evidence, such as past achievements and feedback from others to maintain, enhance, protect, or restore self-esteem. Figure 7.3 presents the self-esteem specific self-report response processes model when episodic retrospective or prospective autobiographical retrieval is involved. When episodic declarative autobiographical retrieval is used, the self-regulation processes theoretically are not involved.

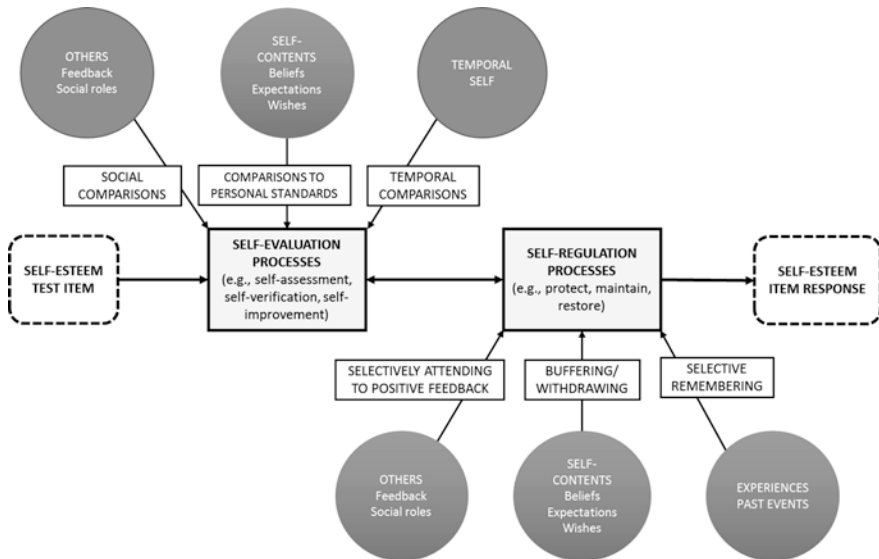


Fig. 7.3 Theoretically expected self-esteem self-report response processes

## ***Study Description***

We recruited 30 adults (18 women, 12 men) ages 20–56 from the general community using posters, announcements to university classes and community centres, Craigslist, Facebook, and word-of-mouth. Participants identified themselves as Caucasian ( $n = 15$ ) or South-East Asian (i.e., Filipino, Vietnamese, Chinese;  $n = 15$ ). All participants had a high school degree or higher education.

We used cognitive interviewing and specifically, think aloud protocols (TAP) and verbal probing (VP), to explore the response processes used while respondents completed the RSES. Cognitive interviewing is a broad class of methods that aim to understand the mental processes employed by respondents when responding to items (Willis, 2005). While completing the RSES, participants were asked to think aloud (i.e., concurrent verbalization TAP). We also asked respondents to clarify their statements, define key terms in their own language, or probed their cognitive, motivational, and emotional processes as they completed the measure (i.e., VP using the Survey Interaction Model; Jobe, 2003). Data collection was conducted by the first author. Each participant's session was digitally audio-recorded and later transcribed verbatim by a professional transcriptionist.

In analysing the transcripts, we examined the response processes and content associated with participants' TAP and VP responses using a theoretically informed coding scheme based on the response processes model. We also reviewed the transcripts to identify contents and themes that had not been included in the coding scheme (i.e., open coding).

## ***Results***

**Observed Responses** In responding to the RSES items, respondents evaluated themselves in terms of worth, competency, and self-liking. Their TAP and VP responses illustrated instances of achievements or positive reflected self-evaluations (e.g., in response to item 5, one participant explained that "... *my family, my job, you know, all the work I do with people around me so I feel like I've got lots to be proud of*"), shortcomings and failures, as well as anxiety about anticipated challenges that could negatively impact their self-esteem (e.g., in response to item 10, one participant noted "*Sometimes I lack confidence. I feel I could not do it, yeah, I'm afraid that I will fail*").

**Content** In the current sample, Caucasian respondents consistently reflected the main content and dimensions of the self-esteem construct (e.g., competence, self-liking, and self-worthiness). All Caucasian respondents referred, in their answers, to most of the main aspects of the self-esteem construct as defined by the theory of self-esteem that influenced the development of the RSES (Rosenberg, 1965). The most references were to competence and a sense of mastery (e.g., in responding to item 5, one participant stated "*I'm going to again go with an 'agree' there, mainly*

*because of my past and what I've come back from. So, I was quite low at one point in time. I was quite depressed...quite upset at a lot of things, and I have been going back in the right direction for some time now, but I haven't made it all the way back, so...I'm going to put a 3 there*"), followed by references to relational self-worth, and to self-respect. In addition to the main dimensions of the self-esteem construct as depicted in the RSES, a few participants talked primarily about meaning in life and future plans, compassion and mindfulness, well-being and quality of life, and spirituality when answering the RSES items. These responses may raise issues regarding the construct representation of some of the RSES items and the potential for conflating theoretical dimensions related to self-esteem with those relevant to other constructs (e.g., meaning in life).

South-East Asian participants' responses reflected content that was not consistent with the theory of self-esteem construct depicted by the RSES. One participant indicated that self-esteem does not exist as a construct whereas other participants referred to self-respect and self-worth in ways that are not consistent with the dominant content of the North-American culture (e.g., self-respect was defined primarily as respecting others).

*Response Strategies* Test takers' observed responses were aided by using a variety of strategies or approaches, such as social comparisons, self comparisons, temporal comparisons, selective remembering, buffering or compensation, normalizing, and counterfactual thinking (i.e., thoughts of what might have been or alternative pasts). For example, one participant noted, when responding to item 4, "*People within my workplace certainly respect me for what I'm doing and what I know and they would come to me so I must be as able as them or more able because they're coming to me to consult with me*" (i.e., social comparisons). Another participant, when responding to item 2 said, "*this is normal, all humans feel not good at all at times, it's human nature; we are not perfect*" (i.e., normalizing).

**Response Processes** Self-referential processing, together with self-evaluation and self-regulation processes, were the main categories of response processes mobilized by respondents when responding to the RSES items.

*Self-Referential Processing* All participants engaged predominantly in self-referential responding, either spontaneously or when prompted by the interviewer's probes. The prevalence of self-referential processing was evident in the following instances: the preponderance of first person reporting, the sustained access to self-characteristics and autobiographical events, and the spontaneous propensity to compare and contrast self and others (i.e., intersubjectivity). These findings suggest that the RSES represents a measure apt to engage self-referential processing, as expected while completing a self-report measure. A corollary of this is that RSES scores can be interpreted as pertaining to self-characteristics, and involve self-processing. However, the depth and breadth of self-referential processing varied greatly across the items and it seemed to be significantly impacted by some very general or abstract items that predisposed some respondents to a more semantic way of processing and reporting information about self (e.g., mobilizing beliefs and implicit theories about self). Reflecting on this, one of the respondents stated:

Participant: *“It’s kind of like a very quick assessment based on my memories – I draw upon them a lot and, uh, I guess I do have a...somewhere in my mind, a compartment of, you know, how I feel about myself, and certain things I’ve decided upon as well, not just...”*

Interviewer: *“So you pooled information from there, because you kind of already decided?”*

Participant: *“Yeah, yeah, I already knew how I felt about myself. So yeah I already decided and just drew on those memories.”*

This tendency to resort to semantic processing increases, as expected, with the degree of generality and abstraction of the items. Here is one pertinent comment made by a participant:

Participant: *“I feel solid in those items that are asking, you know, sort of overall things. Then I went straight to what I already knew and already decided about myself. The questions that are more vague, I would maybe answer that differently at a different time depending on what will be on my mind at that time. But these general, overall things will never change: I already decided these, you know, overall things about myself.”*

Interviewer: *“Could you tell me what are these overall things about yourself that you feel that will not change?”*

Participant: *“I am a person of worth, for example. And I can do things better than others, I am a good, competent worker. I am a kind person. I help people. I like myself. These kinds of things. I mean, I know who I am, right?”*

*Self-Evaluation Processes* Self-evaluation processes supported by a mix of cognitive and affective sub-processes were the most frequent response processes that were evident but they were differentially involved across items. Whereas positively worded items (items 1, 3, 4, 7 and 10) engaged mainly participants’ self-evaluative response processes in light of stable, trait-like personality characteristics (i.e., declarative self-knowledge), the negatively worded items (items 2, 5, 6, 8 and 9) primarily activated respondents’ self-evaluation processes based on episodic retrieval (e.g., experiences of personal failure, incompetence, or rejection).

Moreover, items 2, 5, 6, and 9 mobilized predominantly participants’ self-critical evaluation processes (i.e., measuring oneself against others’ or one’s own expectations, falling short of these expectations, or struggling to meet expectations). For example, in responding to item 2, one participant stated *“Well here you are always competing with people. And sometimes you are above, and sometimes you are on the bottom and, and in those times I, I say no, I am not good at all, sometimes you discourage yourself and then you realize that you, well at that time you, you were at the bottom. In comparisons with others, always.”* Items 1, 3, 8 and 10 predominantly engaged agentic (i.e., intentional) self-evaluation processes in terms of moral values, personal integrity, and self-transcendence. For example, in responding to item 1, one participant said *“Everything is good with my moral compass, I am satisfied with myself”*. It is noteworthy that most of the items that predisposed participants to



engage in reflective-agentive self-evaluation processes were the same items that solicited prospective, constructive, or anticipatory response processes, which suggests that they may be part of a more common generic response process or mechanism.

While responding to the RSES items, all participants engaged in the following self-evaluative strategies in order to select responses to the RSES items: (a) accessing objective self-evaluative information (e.g., test results or achievements), (b) temporal comparisons, (c) downward and upward social comparisons, and (d) searching for, or reporting, evaluative feedback from others. Overall, a consistent engagement in any or all of the four fundamental self-evaluative processes (i.e., self-assessment, self-enhancement, self-verification, or self-improvement) was discernible in participants' responses. Therefore, these results support the interpretation of the RSES scores as self-evaluative judgements.

*Self-Regulation Processes* When responding to the RSES items, the respondents mobilized a multitude of self-regulation processes such as: creating self-esteem (i.e., searching for approval, engaging in value-consistent behaviours), conceptualizing self-esteem (e.g., developing personal theories based on autobiographical data; organizing the facts in a way congruent with positive self-representations), maintaining and managing self-esteem (e.g., reducing dissonance, refusing to engage in behaviours that contradict one's already established self-esteem or personal values, rationalizations, normalization, approval seeking, or success seeking), avoiding further loss of self-esteem (e.g., avoiding challenging situations, over-compensating, complying), enhancing self-esteem (e.g., self-talk, self-serving bias, emphasis on future possibilities of self-improvement), and restoring self-esteem (e.g., re-engaging in socially sanctioned, value-consistent, and worthy behaviors; self-serving judgments; meaning making in response to threats to self-esteem). In particular, agreement with negatively worded items (items 2, 5, 6, 8 and 9) led to a plethora of self-regulation strategies meant to protect or enhance one's self-evaluations (e.g., restoring, compensating, normalizing, or maintaining self-esteem levels). The present findings support the hypothesis that self-regulation processes represent one of the theoretically expected underlying mechanisms of the RSES scores.

*Cognitive-Affective Self-Related Processes* When responding to the RSES items, participants engaged in self-related cognitive processes such as: autobiographical retrieval or episodic memory processes, declarative retrieval pertaining to self-knowledge, counterfactual thinking, and self-attribution processes. In addition, respondents reported experiencing or remembering self-conscious emotions, such as anger, embarrassment, shame, and pride. For example, one participant explained "I'm going to put 'agree' only because I'm angry at myself for past stuff and failures". All of these response processes are consistent with theoretically expected response processes of self-report self-esteem items and, thus, support the inferences made from RSES items.

*Constructive or Anticipatory Response Processes* The constructive-prospective dimensions of the proposed model were reflected in respondents' frequent references to future goals, desires, and imagined possibilities. Their choice of responses to RSES items appeared to be shaped by these intentional goal oriented and future focused mental processes (e.g., one participant stated "I feel that I could become a



*better self. I have to. I want more from myself. I want to be better in the future. So I disagree*” in response to item 1). Therefore, an accurate interpretation of test scores should refer not only to participants’ remembered experiences and beliefs about their self-esteem but also to their future goals, hopes, wishes, and expectations as they shape a positive view of self. RSES scores appeared to provide an answer not only to ‘what is my self-esteem based on my past and present experiences?’ but also to the question ‘what could my self-esteem become?’. In this sense, the RSES scores allowed for a dynamic form of assessment of self-esteem situated in the “zone of proximal development” (Vygotski, 1994, p. 53) or ‘proximal self-esteem’.

## ***Discussion***

We will briefly discuss these findings using the framework of the response processes model of self-report self-esteem items (Figs. 7.2 and 7.3). Overall, when responding to the RSES items, participants demonstrated most of the theoretically expected response processes included in the proposed model. Specifically, participants’ answers reflected response processes consistent with self-referential, self-evaluative, and self-regulation response processes. Participants also engaged in cognitive-affective self-referential processing. The ratio between episodic and semantic retrieval components was, at times, in favour of semantic retrieval for some items, which suggests that some of the RSES items tended to evoke general, abstract evaluations of self.

In terms of content, when responding to the RSES items, the Caucasian respondents reported response content consistent with the construct theory such as: competence, values, self-worthiness and likability, social role expectations and identifications, as well as needs, motives, and expectations. However, the content of South-East Asian participants’ responses reflected content and meanings that were not included in the construct theory underlying the RSES (e.g., openness to negative feedback). This finding raises significant concerns about using the RSES with culturally diverse populations as well as cautions about the validity of inferences made from RSES scores when the measure is used in non-North American (or perhaps non-Western) cultures.

## **Refining the Response Processes Model for Self-Report Self-Esteem Items**

### ***Response Processes***

In conducting this study, we identified several relevant response processes that were not included in the preliminary model. In this section, we will propose a way of integrating these processes into a revised response processes model of self-report,

**Table 7.1** Response processes not anticipated in the original self-esteem self-report response processes model

Response Processes	Observed Behaviour or Expressed Mental Activity
Reflective-agentic evaluative processes	Reflecting upon one's values and general human values
	Making choices consistent with these values
Self-transcendence processes	Contributing to something or someone beyond the self
	Making a difference beyond one's self-interests
	Harmonizing self-interests with communal interests
Additional self-regulation processes	Normalizing
	Self-talk
	Reshaping/recalibrating the meaning of some items (e.g., "not no good but not good enough"; "not useless but helpless")
	Changing the response format for the Likert-type scale (e.g., wanting to select '2.5' instead of '2' or '3')
Cognitive processes	Counterfactual thinking
Embodied responses	"Gut feeling"
Interpersonal, dialogical processes	Struggling to come up with an answer and asking the interviewer for validation; taking items very seriously, dialoguing back and forth to decide on a final answer

self-esteem items. Self-evaluation is not only a self-critical evaluation with respect to standards and expectations set by self or others, but also a reflective-agentic evaluation in light of values, purposes, and self-transcendence (e.g., "good for"). Whereas the theoretically expected response processes model emphasized the self-critical evaluative processes, the revised model will include reflective-agentic self-evaluative processes as complementary processes.

Examples of self-regulation processes that were not initially anticipated but should be included in the revised model are: recalibrating the (response) scale of a measure, reshaping the meaning of the construct or words used to describe it, impression management, use of self-serving bias, positive self-talk, and counterfactual thinking (see Table 7.1). The primary mechanisms underlying responses to the RSES items (i.e., self-evaluation and self-regulation) engage auxiliary affective processes (e.g., self-conscious emotionality), motivational-teleological processes (e.g., intentionality), and cognitive processes (e.g. autobiographical episodic memory and declarative self-knowledge). These auxiliary processes are actively involved in making decisions about test scores when answering the RSES.

### *Content and Dimensions of the Self-Esteem Construct*

The results of the data analyzed in this study support the existence of several components of the self-esteem construct: (a) self-competence ("good at"), (b) self-worthiness ("being intrinsically good"), (c) self-liking ("feeling good" about self), and (d) self-transcendence ("good for"). For the South-East Asian respondents,

“being good with” in the sense of being able to harmonize with others may represent a significant dimension of self-esteem stemming from collectivistic socio-cultural stances. This suggests that, conceptually, self-esteem may not be a unidimensional construct and that respondents may refer to several distinct dimensions of self-esteem when answering self-report self-esteem items. Whereas previous research studies distinguished between competence and self-liking as separate aspects of self-esteem, the present study suggests that intrinsic self-worthiness and self-transcendence may also be important construct-relevant facets of self-esteem. Future research studies will need to evaluate this hypothesis.

### ***Explaining Observed Test Score Variation***

Based on the current findings, we hypothesize that relatively low scores on the RSES may be the result of a combination of the following processes connected with self-evaluative and self-regulation capacities: (a) ineffective regulation processes meant to restore or maintain self-esteem, (b) intense self-conscious emotionality overwhelms the self-evaluation (e.g., feeling based decisions), (c) over-active self-improvement micro-processes during self-evaluation lead to self-dissatisfaction, (d) overwhelming, unprocessed, unintegrated, and highly invested autobiographical events that render self-regulation ineffective and distort self-evaluation processes, (e) the exclusive presence of critical self-evaluation via expectations and values at the expense of reflective agentic self-evaluation (contribution), or (f) the lack, paucity, or inaccessibility of solid sources of self-validation. It is important to note that moderately low scores for South-East Asian participants may not be indicative of moderately low self-esteem because of how self-esteem is culturally defined. Therefore, although using the same self-evaluation and self-regulation mechanisms, South-East Asian participants may recruit different contents to support their self-evaluations.

Relatively high scores on the RSES may be due to: (a) strong self-regulation processes able to offset any perceived threat to self during triggering “self-esteem moments”, (b) the predominance of reflective self-evaluation, (c) the variety, richness, and heightened accessibility of self-esteem sources, and (d) the predominance of situationally negative self-attribution and positive stable self-attributions. Very high scores on the RSES may be due to self-evaluative processes that rely primarily on ready-made beliefs shaped by normative social messages (e.g., one has to be positive, one should have high self-esteem), and to self-regulation processes related to impression management and socially desirable responding.

It is critically important to evaluate to what extent the self-esteem self-report response processes model can be applied cross-culturally. Although it is possible that the core response processes (i.e., self-evaluation and self-regulation) may be culturally invariant, how exactly they are implemented and what contents they recruit may be significantly different across cultures. Therefore, a contextualized model of responding to the RSES items would be crucial in order to provide accurate interpretations from the test scores.

## Some Concluding Comments

### *Response Processes Evidence with Respect to Inferences Based on RSES Scores*

Current findings generally support that the RSES is a traditional self-report measure of trait global self-esteem that assesses participants' stable evaluative stances towards self that are mainly grounded in general and already established (ready-made) beliefs about self. RSES items tend to be acontextual, non-specific, abstract and general, and therefore, make it difficult for respondents to retrieve specific personal experiences and the emotional tone associated with those experiences. Qualifiers such as: "overall", "on the whole", and "all in all" appear to have pulled respondents into generalization processes and towards a very abstract level of processing and formulating or accessing implicit theories about self and others somewhat disconnected from personal experiences. Therefore, although the RSES may be very helpful in providing information about respondents' general and decontextualized beliefs about their competence and self-worth, it may be largely unsuitable for tracking progress and formulating or evaluating interventions related to self-esteem.

In spite of being a self-report measure, many times semantic and abstract processing and responding took precedence over episodic, affective responding. Some rather emotionally laden and highly evocative wording ("no good at all", "failure", "useless") and some radical or extreme item wording (e.g., "All in all I am a failure") seemed to be effective in triggering "self-esteem moments" and the associated response processes, and this type of wording unexpectedly compensated for the highly abstract and acontextual nature of the scale by eliciting a multitude of emotional and motivational processes meant to protect or enhance threatened self-esteem. The content of responses appropriately matched the main theoretical dimensions of the self-esteem construct defined as a mixture of competence and self-worth. However, it is critical to note that this is accurate only for the Caucasian respondents, and not for the South-East Asian respondents.

### *Building and Using a Model of Response Processes as a Source of Validity Evidence*

To our knowledge, this study represents the first investigation of the substantive validity of inferences from RSES scores using a response processes inquiry. Given the widespread use of this scale, these results are very important given that they help clarify some important aspects related to how to interpret RSES scores in future research and in clinical practice. Our main focus in this chapter was to present a self-report response processes model and apply it to the RSES as an example or 'proof-of-concept'. In conducting a response processes study as a source of validity

evidence however, we would recommend that validity evidence be examined and presented in greater detail on an item-by-item basis. Like test content as a source of validity evidence, this kind of an analysis could be particularly useful during test development.

Being the first of its kind, this study is highly exploratory in its nature. The composition of our sample was limited in terms of demographic diversity such as ethnicity (although the obtained groups led to some important observations about how self-esteem was conceptualized by each group) or even age and gender. Hence, the findings of this study may only be applicable to similar samples, and no conclusions should be extended beyond the context of this study without additional work to replicate these findings and expand this type of research using different samples and measures. Moreover, the method used to investigate response processes (i.e., TAP combined with VP) may not be the most conducive for exploring the situated and socio-cultural dimensions of a response processes model. It would be helpful if future research studies could implement a methodology that could intentionally target the socially situated response processes. Finally, our model makes reference to the speed of processing, with episodic autobiographical retrieval requiring longer processing times than declarative autobiographical retrieval. We did not examine if speed of processing differed depending on the type of retrieval, although this is something we recommend be examined in future studies.

We hope this study may serve as an example for future empirical investigations using response processes as a source of validity evidence. Our aim was to encourage and inspire future research by contributing to the empirical findings regarding the validity of inferences made from RSES scores, but even more so by proposing an example of how to conduct this type of research using a response processes model. This research project should be viewed as laying the foundation for future research in the sense of providing a self-report model of response processes that can be studied further and applied to other self-referential constructs examined using self-report measures or items.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baumeister, R. F., & Vohs, K. D. (2004). *Handbook of self-regulation: Research, theory, and applications*. New York, NY: Guilford Press.
- Blascovich, J., & Tomaka, J. (1993). Measures of self-esteem. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 57–71). Ann Arbor, MI: Institute for Social Research.
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.
- Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, 53, 594–628. <http://dx.doi.org/10.1016/j.jml.2005.08.005>

- Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.
- Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, *30*, 161–172. doi:[10.1177/0146167203259930](https://doi.org/10.1177/0146167203259930).
- Hoyle, R. H., & Wiley-Blackwell Online Books. (2013). *Handbook of personality and self-regulation* (1st ed.). Chichester, UK/Malden, MA: Wiley-Blackwell.
- Jobe, J. B. (2003). Cognitive psychology and self-reports: Models and methods. *Quality of Life Research*, *12*, 219–227.
- Kim, H. (2012). A dual-subsystem model of the brain's default network: Self-referential processing, memory retrieval processes, and autobiographical memory retrieval. *NeuroImage*, *61*, 966–977. doi:[10.1016/j.neuroimage.2012.03.025](https://doi.org/10.1016/j.neuroimage.2012.03.025).
- Kurczek, J., Wechsler, E., Ahuja, S., Jensen, U., Cohen, N. J., Tranel, D., & Duff, M. (2015). Differential contributions of hippocampus and medial prefrontal cortex to self-projection and self-referential processing. *Neuropsychologia*, *73*, 116–126. doi:[10.1016/j.neuropsychologia.2015.05.002](https://doi.org/10.1016/j.neuropsychologia.2015.05.002).
- Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. *NeuroImage*, *133*, 151–162. doi:[10.1016/j.neuroimage.2016.02.014](https://doi.org/10.1016/j.neuroimage.2016.02.014).
- Levine, B., Turner, G. R., Tisserand, D. J., Graham, S. I., Hevenor, S. J., & McIntosh, A. R. (2004). The functional neuroanatomy of episodic and semantic autobiographical remembering: A prospective study. *Journal of Cognitive Neuroscience*, *16*, 1633–1646.
- Magno, E., & Allan, K. (2007). Self-reference during explicit memory retrieval. An event-related potential analysis. *Psychological Science*, *18*, 672–677.
- Mu, Y., & Han, S. (2010). Neural oscillations involved in self-referential processing. *NeuroImage*, *53*, 757–768. doi:[10.1016/j.neuroimage.2010.07.008](https://doi.org/10.1016/j.neuroimage.2010.07.008).
- Northoff, G., & Bermphol, K. (2004). Cortical midline structures and the self. *Trends in Cognitive Sciences*, *8*, 102–107.
- Northoff, G., Heinzel, A., de Greck, M., Bermphohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain – A meta-analysis of imaging studies on the self. *NeuroImage*, *31*, 440–457.
- Rahamim, O., Garbi, D., Shahar, G., & Meiran, N. (2016). Evaluative processes in self-critical individuals: The role of success and failure inductions. *Personality and Individual Differences*, *100*, 105–113. doi:[10.1016/j.paid.2016.03.083](https://doi.org/10.1016/j.paid.2016.03.083).
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Sajonz, B., Kahnt, T., Margulies, D. S., Park, S. Q., Wittmann, A., Stoy, M., & Bermphohl, F. (2010). Delineating self-referential processing from episodic memory retrieval: Common and dissociable networks. *NeuroImage*, *50*, 1606–1617. doi:[10.1016/j.neuroimage.2010.01.087](https://doi.org/10.1016/j.neuroimage.2010.01.087).
- Somerville, L. H., Kelley, W. M., & Heatherton, T. F. (2010). Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. *Cerebral Cortex*, *20*, 3005–3013. doi:[10.1093/cercor/bhq049](https://doi.org/10.1093/cercor/bhq049)
- Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, *21*, 489–510. doi:[10.1162/jocn.2008.21029](https://doi.org/10.1162/jocn.2008.21029).
- Sui, J., & Humphreys, G. W. (2013). Self-referential processing is distinct from semantic elaboration: Evidence from long-term memory effects in a patient with amnesia and semantic impairments. *Neuropsychologia*, *51*, 2663–2673. doi:[10.1016/j.neuropsychologia.2013.07.025](https://doi.org/10.1016/j.neuropsychologia.2013.07.025).
- Taylor, S. E., Neter, E., & Wayment, H. A. (1995). Self-evaluation processes. *Personality and Social Psychology Bulletin*, *21*, 1278–1287.
- Vygotsky, L. S. (1994). The problem of the environment. In R. van der Veer & J. Valsiner (Eds.), *The Vygotsky reader* (pp. 338–354). Oxford, UK: Blackwell.

- Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.
- Zeigler-Hill, V., & Ebooks Corporation. (2013). *Self-esteem*. New York, NY: Psychology Press. doi:[10.4324/9780203587874](https://doi.org/10.4324/9780203587874).
- Zell, E., & Alicke, M. D. (2009). Self-evaluative effects of temporal and social comparison. *Journal of Experimental Social Psychology*, *45*, 223–227. doi:[10.1016/j.jesp.2008.09.007](https://doi.org/10.1016/j.jesp.2008.09.007).
- 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, NC: Information Age Publishing.

# Chapter 8

## Response Processes and Validity Evidence: Controlling for Emotions in Think Aloud Interviews

Jacqueline P. Leighton, Wei Tang, and Qi Guo

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) are clear about the importance of response processing data as a source of validity evidence for test and item score interpretations. For example, the response processes used by test-takers of different ability levels can be used to confirm expectations about the cognitive or information-processing strategies underwriting correct and incorrect responses to test items. Aside from serving as evidence for validity arguments, these data can also inform test item design and development (Leighton & Gierl, 2007) by revealing construct-irrelevant aspects of tasks that test-takers may fail to understand and thus impede performance.

However, despite at least 25 years of attempts by psychometric practitioners and researchers to better integrate response processes into validation efforts, key variables have not been fully considered for gathering and interpreting these data. In particular, one class of variables that has remained virtually unexplored in terms of its effect on response processing data is examinees' affective or emotional states. Considering these states may be important, especially in think-aloud interview studies (see Ericsson & Simon, 1993 for think-aloud interviews and protocol analysis), they require investigation. For example, there is reason to suspect that evaluative anxiety may disrupt aspects of response articulation (i.e., expressing problem solving processes) and the actual problem solving engaged in by test-takers in think aloud interviews, thus undermining the integrity of the data (Leighton, 2013; Norris, 1990; see also Hsu, Babeva, Feng, Hummer, & Davison, 2014, for effect of distractions on response processing data). This would be especially problematic given that

---

J.P. Leighton (✉) • W. Tang • Q. Guo

Center for Research in Applied Measurement and Evaluation (CRAME), Department of Educational Psychology, Faculty of Education, University of Alberta,  
6-119D Education North Building, Edmonton, AB T6G 2G5, Canada  
e-mail: [jacqueline.leighton@ualberta.ca](mailto:jacqueline.leighton@ualberta.ca); [wtang3@ualberta.ca](mailto:wtang3@ualberta.ca); [qig@ualberta.ca](mailto:qig@ualberta.ca)

© Springer International Publishing AG 2017

B.D. Zumbo, A.M. Hubleby (eds.), *Understanding and Investigating Response Processes in Validation Research*, Social Indicators Research Series 69,  
DOI 10.1007/978-3-319-56129-5\_8

137



response processing data are designed to inform us about how well individual test items measure specific cognitive or information-processing skills.

Although understanding the relationship between test-takers' emotional states and their response processes may be necessary to accurately inform validation studies, this relationship will undoubtedly be complicated. First, affect and emotion are defined in various ways and, most often than not, in overlapping ways (see Shuman & Scherer, 2014). Clear operational definitions of these terms are essential for further exploration. Second, affect and/or emotions may impact response processing data in think-aloud interviews, for example, in ways that are distinct from the ways in which they influence actual test-taking behavior. Thus, considering the results of think-aloud interviews and real test-taking situations is an important consideration when accounting for emotional states in response processing data. Third, some emotions may prove to be more disruptive than others. For example, evaluative or test anxiety, which is commonly observed in testing, is the most extensively studied academic emotion (Pekrun, Goetz, Titz, & Perry, 2002) and a thorough exploration of its effects may be a priority when accounting for emotions in response processing data.

In the balance of this chapter, we attempt to make inroads by distinguishing overlapping concepts such as affect and emotion. In particular, we provide a brief review of emotions from a dynamic perspective and identify the relevance of emotions with a focus on evaluative, test anxiety in academic performance and by extension response processing data. We summarize recent think-aloud studies suggesting a link between evaluative anxiety and its potential to degrade response processing data. We conclude by discussing the need to account and control for emotional variables in think-aloud studies, and a call for research to address the impact of emotions in the collection and interpretation of response processing data for the purpose of validation.

## **The Construct of Emotion**

### ***Emotion Versus Affect***

A review of the literature suggests that emotion is a construct that builds on multiple variables, including affect. For example, a recent and comprehensive definition of emotion indicates it is a multifaceted phenomenon, comprising affective, cognitive, physiological, motivational, and expressive components (Kleinginna & Kleinginna, 1981; Shuman & Scherer, 2014). Each component is further specified by its valence (i.e., positive versus negative), arousal (i.e., activated versus deactivated) and intensity (i.e., strong versus weak). Affect is often considered to be the most salient building block or component in emotion. However, in comparison to emotion, it is amorphous and denotes only the unconscious, general experience attached to a feeling or sensation; this feeling is often accompanied by bodily or physiological

changes (Russell, 2003). Affect is often used to denote a broad variety of unnamed positive and negative states (Pekrun & Linnenbrink-Garcia, 2014). For example, affective states can involve highly changeable experiences or what have been called “event dependent” feelings (see Shuman & Scherer, 2014; Turner, Christensen, Kackar-Cam, Trucano, & Fulmer, 2014). To be considered an emotion, affect has to be recognized and named.

Thus, the cognitive component of emotion is also important as it involves appraisal, that is, naming or self-evaluating the feeling that is being experienced (see Moors, Ellsworth, Scherer, & Frijda, 2013; Scherer, Schorr, & Johnstone, 2001). The cognitive component may also involve reflections or judgments about the experience (Russell, 2003; see also Mayer & Gaschke, 1988), and can play a special role in triggering additional sensations (see Sander, Grandjean, & Scherer, 2005; Scherer et al., 2001).

The physiological component refers to bodily changes or arousal. For example, the emotion of anxiety can manifest itself consciously with concomitant physiological changes such as increased heart rate, sweat gland and bladder activity. Moreover, anxiety can present an elevated neuroendocrine response “consist[ing] of an increase in epinephrine and norepinephrine, cortisol, growth hormone and prolactin” (Hoehn-Saric & McLeod, 2000, p. 217). The motivational component in emotion is often manifested in the volition of different behaviors (Maehr & Mayer, 1997), and the expressive component serves to externalize inner feelings in tangible ways such as in the demonstration of a facial expression or other physical action (Pekrun & Linnenbrink-Garcia, 2014).

To understand how these components work in unison, consider the following example: A student judges a test as too demanding without enough time to complete it (cognitive). Beginning to feel a sense of unease (affective), his or her pulse starts to rise and heart rate increases (physiological). The student may start to answer questions with heightened attention (motivation), unconsciously tapping his or her foot or tightening the grip on the pencil (expressive). Then, a series of judgments (cognitive) may unfold – the student perceives the unpleasant experience, acknowledges what is being felt as anxiety and fears that this state will influence performance on the test.

### ***Emotion Regulation***

The construct of emotion is described in modern theories as a multifaceted, interactive process (i.e., affective, cognitive, physiological, motivational and expressive components) (see Kleinginna & Kleinginna, 1981; Shuman & Scherer, 2014). Changes in one component can influence other components and thus alter the course emotion takes. Consider again test anxiety: A student thinks about failing an exam (cognitive), which leads to feelings of unease (affective) and an increase in heart rate (physiological). If the feelings of unease and increased heart rate continue without regulation, the student may begin to experience an intensity in these feelings

(affective, physiological), and stop responding to the test (expressive). Stopping the test could lead the student to bolster the feelings of unease (affect) already being experienced. Alternatively, stopping the test could reduce the unease if the student uses the time to relax (expressive, physiological) and thus lessen the unpleasant feelings.

Although the interaction among components of emotion is complex, patterns in emotions may be observed in certain activities. Goldin (2000) calls these patterns *emotional pathways* and indicates that within particular events, for example, the interplay among affective, cognitive and motivational components may be anticipated. Toward this end, in a qualitative study, McCulloch (2011) found that high school calculus students using graphing calculators experienced a specific sequence of emotions – from frustration to curiosity, then discouragement, helplessness or annoyance, and finally embarrassment. A benefit of detecting emotional patterns is that they can be anticipated and thus regulated in positive ways. Not surprisingly, the cognitive component is key in this respect. In particular, meta-cognitive functions, including self-monitoring (e.g., “I know exactly how I’m feeling”) and self-evaluation (e.g., “It is not right that I’m feeling this way given how much I studied”) can serve to detect, interpret and respond to affective, physiological, motivational and other bodily changes. Constructive self-talk and thoughts that lead to changes in behavior can feed into the interactive emotional process and guide positively reinforcing responses to deal with the events (see Malmivuori, 2006).

## Relevance of Emotions in Educational Achievement Testing

Students experience a variety of emotions in academic settings that influence their learning and achievement (Pekrun, 1992; Pekrun et al., 2002). For example, in a qualitative study, Pekrun et al. found students reported a consistent set of emotions while engaged in class, studying, and taking tests and other assessment activities. Although anxiety was identified as the single most often reported emotion, they found positive emotions such as enjoyment, hope, pride, and relief, in addition to negative emotions such as anger, boredom, shame, and hopelessness. Less frequently reported emotions included gratitude, admiration, contempt, and envy. It is only recently that the range of emotions students experience at school has begun to receive the full attention it deserves (Pekrun et al., 2002). Over a decade ago, Pekrun et al. (2002, p. 91) wrote “Academic emotions have largely been neglected by educational psychology, with the exception of test anxiety.” To be sure, since 2002, this research gap has begun to be addressed, and as an example, the first *International Handbook of Emotions in Education* was released in 2014, including chapters on anxiety, interest, enjoyment, boredom, shame and pride (see Pekrun & Linnenbrink-Garcia, 2014). Moreover, the study of a wider range of emotions than just anxiety such as trust and wellbeing have become the focus of interest in educational measurement studies (e.g., Chu, Guo, & Leighton, 2014).

## *Emotion and Cognitive Performance*

Emotions can impact information processing, thus influencing the types of cognitive performances often observed in academic settings. The impact on information processing can be positive or negative, and these effects may be especially pronounced for intensely-experienced emotions (e.g., Carver, Peterson, Follansbee, & Scheier, 1983). For example, boredom impacts cognitive performance primarily by impairing attention (e.g., Pekrun, Goetz, Daniels, Stupnisky, & Perry, 2010); the detrimental consequences of the impairment are likely to grow with increases in the intensity of the boredom. In general, students will have trouble concentrating on tasks when strong negative emotions such as anxiety, boredom, and frustration, as well as positive emotions such as joy, which can also serve to distract attention away from the task at hand. Again intensity appears to be an influential factor. For some emotions, including anxiety and even anger, moderate intensity may be beneficial and facilitate concentration and performance on tasks. However, too much or too little can hinder concentration and thus performance.

The different ways emotions impact cognition can be considered more specifically in terms of the processes of assimilation and accommodation (e.g., Fiedler, 2001; Fiedler & Beier, 2014). According to Piaget (1954), assimilation and accommodation are complementary processes in how human beings make sense of incoming information in light of what is already known (also known as cognitive adaptation). Assimilation is the process by which new, incoming information is understood based on previously held beliefs even at the cost of altering – and possibly misrepresenting – the actual input of information. Accommodation is the process by which new, incoming information is understood by adjusting previously held beliefs so as to properly represent the new information.

Emotions influence assimilation and accommodation in distinct ways (e.g., Forgas, 1998; Sinclair & Mark, 1995). For example, negative emotions such as depression have been found to facilitate accommodation by prolonging stimulus processing (Forgas, 1998), avoidance of careless mistakes (Sinclair & Mark, 1995), and generating concrete and detailed representations or understanding of information (Beukeboom, & Semin, 2006). In contrast, positive emotions such as joy have been found to facilitate assimilation by extending previously held beliefs in the generation of constructive inferences about the new information (Storbeck & Clore, 2005), inducing priming effects and heuristic judgments (Storbeck & Clore, 2008), and flexible representations (Huntsinger, Clore, & Bar-Anan, 2010). Fiedler and Beier (2014) indicate these emotion-specific accommodative and assimilative strategies are self-regulatory. For example, a careful analysis of information, originating from a slightly depressed emotion, can help minimize any mistakes in comprehension and may yield success in task performance; thus, re-establishing positive emotions in the problem-solver. Likewise, loose application of heuristic judgments, originating from a slightly joyful emotion, can lead to mistakes in task performance, introduce negative emotion in the problem-solver as a result, and re-establish a careful, attentive stance in problem-solving.

As applied to response processing data, when students are required to solve test items or are invited to take part in think-aloud interviews that involve test items, their knowledge and skills find expression against a backdrop of different emotions and self-regulatory strategies. Although a range of academic emotions (see Pekrun et al., 2002) has been identified, in the next section we focus the discussion on a particular type of emotion – evaluative or test anxiety. Evaluative anxiety may be the most striking emotion that can negatively impact students’ cognitive performance on educational, achievement tests and, by extension, in think-aloud interviews involving test items.

### *Evaluative Anxiety in Testing and Response Processing*

Despite a general newfound focus on academic emotions, evaluative anxiety, specifically, has long been known to impede student test performance and continues to be a well-established impediment (Cizek & Burg, 2006; Zeidner, 2014). The Yerkes-Dodson law (i.e., the inverted U-shaped curve showing the relationship between performance and arousal; Yerkes & Dodson, 1908) reminds us that mild forms of arousal can be beneficial to performance by focusing attention on the task of interest. However, arousal and anxiety are distinct; arousal is the state of being “alert” or “awake,” whereas anxiety is defined by the American Psychological Association as “an emotion characterized by feelings of tension, worried thoughts and physical changes like increased blood pressure” ([www.apa.org/topics/anxiety/](http://www.apa.org/topics/anxiety/)). Thus, anxiety can be debilitating if experienced in intense forms.

Many students experience anxiety in their academic lives. In fact, according to the American Test Anxieties Association, anxiety is considered to be the dominant source of scholastic impairment with a prevalence rate of 16–20% of students reporting high test anxiety and another 18% reporting moderate anxiety. Anxiety is often most pronounced during performance-oriented school activities such as test taking or in anticipation of an evaluative activity. That test items can provoke anxiety and interfere with examinees’ information and response processing is well established (Cizek & Burg, 2006). This provocation of anxiety is unsurprising given that classroom achievement tests, including large-scale tests such as college-readiness tests, are often high-stakes and gateways to postsecondary opportunities and economic mobility. Testing situations and tasks, which frequently rely on multiple-choice formats and are heavily focused on one-time, snap-shot performances, often arouse concern and even alarm in test-takers (Ryan & Ryan, 2005). For example, many girls and minority test-takers exhibit test anxiety in fear of fulfilling negative stereotypes about their cognitive abilities when completing multiple-choice mathematics tests (Bosson, Haymovitz, & Pintel, 2004; Cohen & Sherman, 2005; Steele, 1997).

Aside from recognizing the harmful effects of anxiety on test performance and trying to provide tips and strategies for reducing anxiety (Cizek & Burg, 2006), there are no formal ways to treat it except to provide test-taking accommodations

for students. Usually these come in the form of providing additional testing time for students. However, to be considered for accommodations under the Americans with Disabilities Act (ADA), students may need to demonstrate evidence of “mental impairment” and show that it substantially limits one or more major life activities (Zuriff, 1997). Having to provide such evidence is not without its social drawbacks. Even if students with debilitating high evaluative anxiety are accommodated, many more undoubtedly experience moderate anxiety, do not report it, and are not accommodated. If so, it begs the question of how such moderate levels of anxiety impact student test performance. Making inferences about non-accommodated students’ achievement test scores in light of moderate anxiety impairments is bound to be challenging given that no formal or systematic procedures exist for recognizing or controlling such impairments.

Evaluative anxiety is usually not taken into account in the interpretation of test scores or in validation efforts with non-accommodated populations; probably due to the assumption that if it is moderate, it is being controlled (i.e., self-regulated) during the test and therefore does not alter overall test performance and score interpretation in a material way. However, in validation studies, where the objective is often to drill down on individual test items designed to measure knowledge and skills, and collect finer-grained data about response processing, moderate anxiety may indeed matter more than one might expect. Moderate anxiety may alter the strategies students use to solve items and therefore influence what test developers conclude about what those items are measuring in students. Surprisingly, few if any published papers or reports outline how evaluative anxiety might influence response processing data gathered in think-aloud interviews – a primary method used to collect response processing data (Leighton, 2004; see also Ericsson & Simon, 1993). Given how much attention is devoted by testing specialists to minimizing error and bias in test score interpretations, surely there is reason to question whether the response processes probed in think-aloud studies are skewed by asking individual students to solve test items aloud while an interviewer observes.

## **Distractions and Disruptions: The Impact of Evaluative Anxiety on Response Processing Data**

Relatively little research has been conducted on the accuracy of response processing data collected using think-aloud interviews in light of the role of evaluative anxiety (e.g., Leighton, 2013; see also Cassady & Johnson, 2002; Zeidner, 2014). Given what is known about the impact of emotions generally on cognitive performance and, specifically, the debilitating effects of test anxiety, response processing data needs to be scrutinized carefully for validating item-based inferences.

Even moderate anxiety may disrupt response processing of test items during think-aloud interviews. First, students participating in think-aloud interviews could perceive the interviewer – perhaps unconsciously – as a judge, serving to provide an

evaluation of students' abilities (and shortcomings), leading to self-consciousness, and associated anxiety. To be sure, even if the think-aloud interview is presented by investigators as low-stakes, with the standard instructions indicating the interview is designed to reveal thought processes without judgment of right or wrong strategies, we have no assurance that students think and act as though this were true. We are not aware of any research to suggest what students might believe about the objectives of the interview (however, see Leighton, 2013, which is discussed later in this chapter). In fact, given that the interviewer is observing how participating students solve the items or tasks, probes them to "keep talking" every 15 s or so, the situation may present, at face value, as relatively nerve-wracking for students (Leighton, 2013; see also Kyllonen, *in press*; Sawyer & Hollis-Sawyer, 2005; Steele, 1997). In other words, simply engaging with test items in front of an investigator or interviewer may elevate stress levels for students. Second, and related to the first, assuming that students perceive the think-aloud interview as another vehicle for the assessment of their abilities, evaluative anxiety may impact their performance in perhaps worse ways than under usual testing circumstances.

### *Distractions and Disruptions to Cognitive Processing*

The cognitive mechanism by which anxiety is expected to influence response processing involves impairment of working memory or its central executive. For example, a dominant view has been that the anxiety experienced during evaluative situations leads to excessive self-monitoring (regulation) of performance, which overburdens working memory and therefore leaves few attentional resources to focus on the actual task (see Sarason, Sarason & Pierce, 1995). Beilock, Kulp, Holt, and Carr (2004) identified a state of "choking under pressure" in situations where performance is considered to be highly valued and potentially used to make ego-threatening inferences about individual achievement or intellect (i.e., "Am I smart enough?"). Students in these performance-oriented situations have been found to engage in excessive self-monitoring, the result of which is disrupted response processing because working memory is tied up and, thus, leads to suboptimal outcomes (see also Beilock & Carr, 2001; Ericsson, 2006; Lewis & Linder, 1997). In general, excessive self-focus reduces attention on problem solving and impairs performance.

Another more recent mechanism is outlined in attention-control theory (Eysenck, Derakshan, Santos, & Calvo, 2007). In attention-control theory, anxiety is shown to disrupt performance not necessarily because of excessive self-monitoring but because of excessive *outward monitoring*. In other words, the individual is constantly engaging in an outward focused, environmental scan, looking for potential threats; for example, the facial expressions of the interviewer to provide cues about the correctness of the answers. This excessive outward monitoring again hampers the central executive of working memory to regulate and direct needed attentional resources to focus on the task. From a neuroscientific perspective, evidence indicates



that anxiety increases activation of the amygdala, the brain's structure associated with perceiving and controlling emotion, and decreases activation of the prefrontal cortical areas, which are known to be extensively involved in the regulation of attention (Bishop, 2007).

The question of how distractions may influence response processing in think aloud studies was investigated in a recent study by Hsu et al. (2014). They experimentally investigated cognitively-induced distractions (i.e., answering trivia questions, playing a visual puzzle game [tetris], or no distraction [control]) during a think-aloud interview, and found that the distractions significantly altered aspects of response processing. In particular, a content analysis of the verbal reports using the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) revealed students in the two distraction conditions – answering trivia questions and playing tetris – produced a lower number of words during the think aloud session relative to the controls. In addition, answering trivia questions led to the production of more non-fluencies (e.g., “uh” or “umm”) and filler words (e.g., “like” or “you know”) unrelated to the task and greater measured disengagement compared to controls. Although the distractions that Hsu and colleagues examined were not emotionally induced, they still suggest that interferences can reduce the quality of the response processing data (e.g., lower number of relevant words and higher number of filler words). Hsu and colleagues focused on the effects of distraction on aspects of response processing, but they did not employ strict problem solving tasks in the interviews or examine accuracy of task performance.

In an investigation of interviewer variables on students' problem-solving response processes, Leighton (2013) found interviewer domain-knowledge had an effect not only on students' accuracy of performance but also on the quality of their response processes. Leighton randomly assigned 71 high and moderate-ability Grade 12 math students to one of three think-aloud interview conditions (i.e., expert, control, and novice). Each of the conditions was exactly the same in procedure, including task instructions, order of test items, and order of follow-up surveys, except for one variable – a single sentence in the script the interviewer used to describe his or her level of domain knowledge at the start of the interview. This manipulation was intended to influence students' unease and self-monitoring during the interview. In the *novice interviewer condition*, the interviewer introduced himself/herself as a non-expert with a single sentence using the following script:

...before I explain what we will be doing today, let me introduce myself. My name is [X] and I'm from the University of [X]. My area of expertise is not in Mathematics but I've been interested in how students solve problems for many years. So, now let me tell you about the study you're involved with today...

In the *expert interviewer condition*, the interviewer introduced himself/herself as an expert in mathematics and stated *My area of expertise is in Mathematics* – the script was otherwise identical to the novice condition. In the control condition, the interviewer did not include any mention of his or her expertise in mathematics.

Leighton (2013) found that students assigned to the novice condition were more accurate in their task responses for easy, medium and difficult mathematics items



compared to students assigned to the control and expert conditions; suggesting that students in the control condition interpreted the interviewer to be, by default, an expert. (This was confirmed in a post-interview question). Furthermore, students assigned to the novice condition exhibited more sophisticated response processes or cognitive models leading up to their responses (as coded in retrospective verbal reports but not in concurrent verbal reports) on medium and difficult items compared to students in both the expert/control conditions. Although Leighton found no differences across the three conditions on the *Test Attitude Inventory (TAI)* (Spielberger et al., 1980), a self-report measure of state test anxiety, and on measures of familiarity or confidence, and even metacognition (self-monitoring and regulation), differences were found in how students perceived the interviewer. In particular, students in both the expert and control conditions reported in a post-interview survey perceiving the interviewer “as an expert in Mathematics” significantly more often than those students in the novice condition. Additionally, on an indirect and experimental measure of nervousness or unease – frequency of validation seeking speech (e.g., am I on the right path?) – students in the expert and control conditions tended to exhibit more validation seeking speech compared to students in the novice condition, although this difference failed to reach statistical significance (see Hsu et al., 2014 for similar findings).

The results of Leighton (2013) are disconcerting. On the one hand, students in the expert/control conditions should have reported greater state anxiety than students in the novice condition if the manipulation of increasing perceived interviewer knowledge had raised students’ levels of unease. On the other hand, it is possible that the manipulation could have raised students’ unease to a threshold level just below their awareness, thus thwarting measurement using self report, but nonetheless obstructing their cognitive performance. To be sure, self-report measures may be insufficiently sensitive to detect variations in emotional shifts that impact performance (see Kyllonen, *in press*). Toward this end, Berridge and Winkielman (2003) present a growing empirical case for what they call *unconscious emotions*. Unconscious emotions may fail to enter awareness if the underlying feeling or sensation is diffuse, the individual does not know the referent or origin of the feeling, and/or simply mislabels the sensation. An empirical example is borrowed from the work of Zajonc and associates (e.g., Murphy & Zajonc, 1993) who have demonstrated that subliminal affective priming (e.g., smiling or angry faces) can impact preference ratings for an object even though participants are unaware of the priming. Although the most plausible account of students’ reduced accuracy and response processing in Leighton’s (2013) expert/control condition is some form of diffuse affect or unconscious emotion disrupting their attentional focus on test items, this necessitates further research, including the development of tools to measure the emotion in some form during think aloud interviews.

The potential imprecision associated with response processing data gathered from think-aloud interviews has implications for validity arguments. As response processing data are increasingly used to confirm the type, range and sophistication of the response processes examinees use to construct and/or select answers to test items (see AERA, APA, & NCME, 2014), there is reason to question the validity of the data in providing an accurate window into response processing. At the very

least, what we observe in the Leighton (2013) and Hsu et al. (2014) studies is that response processing data may be easily compromised as function of emotional disruptions or cognitive distractions, respectively, activated in response to environmental conditions. If such slight manipulations to the conditions of the think-aloud interview can lead to shifts in response processing and reduce accuracy and quality of reported knowledge and skills, it is necessary to consider the potential of other seemingly benign variables that could be distorting these data.

### ***Approaches for Accounting and Controlling for Emotion***

There are opportunities to investigate the role and potential impact of emotions on the response processing data collected in think aloud interviews, especially those data used for validation purposes. Development of measures of emotions for think-aloud interviews and further exploration of how evaluative anxiety and/or other emotions may influence students' response processes are areas in need of research. Such efforts would be designed to help us improve our procedures in conducting controlled think-aloud studies and thus gather better quality response processing data that are more reflective of test items than participants' reactions to interview materials and conditions. Although controlling for emotional variables is not currently done when collecting response processing data in educational measurement studies, measures of academic emotions nonetheless do exist and may provide a starting point.

**Self-Report Measures** Quantitative self-report measures such as the *Test Attitude Inventory* are the most commonly used tools to evaluate academic emotions; other examples include the Achievement Emotions Questionnaire (Pekrun et al., 2002), and the Epistemic Emotion Scales (Pekrun & Meier, 2011). Quantitative self-report measures have two major advantages – they are practical and standardized. These advantages can also present challenges. For example, most use a Likert-type response format (e.g., 5-point or 7-point rating scale) but these scales can lead to bias originating from individual response styles (Paulhus, 1991) and social desirability effects (Frenzel, 2014). Thus, less obtrusive self-report measures may and should be considered.

*Anchoring Vignettes* and *Forced Choice* tools are designed to reduce the bias originating from response style and social desirability (see King & Wand, 2007; see also Kyllonen & Bertling, 2013). Anchoring vignettes require participants to rate hypothetical examples reflecting intensity levels of a given attribute before rating themselves. For example, consider a vignette for measuring affect – *Ken loves life and is happy all the time. He never worries or gets upset about anything and deals with things as they come* (<http://gking.harvard.edu/vign/eg/affect.shtml>). After reading this vignette, participants consider a set of response categories (e.g., none, mild, moderate, severe, and extreme) and respond to questions such as *overall in the last 30 days, how much of a problem did Ken have with feeling sad, low or depressed?*

Afterwards, participants are asked to rate themselves using the same set of categories and their scores are compared with their vignette ratings to control for their interpretation and use of the response categories. Another way to reduce social desirability is to use forced-choice tools, which require participants to choose responses among sets of options that are equally desirable or undesirable. However, forced-choice tools lead to ipsative scores and are therefore challenging to use to compare across individuals.

Self-reports can also be qualitative and collected using interviews. Questions about students' emotions could, in principle, be collected within a think-aloud interview. However, there is a potential danger with conflating the foci of the think-aloud interview. Think-aloud procedures were designed to measure cognitive or information processes and not emotional responses (Ericsson & Simon, 1993). Probing students, possibly even priming them on felt emotions, during the think-aloud interview could, ironically, bias and further erode the response processing data that are being gathered. Nonetheless, emotions could be probed independently and post-problem solving using other types of structured or semi-structured interviews (e.g., Debellis & Goldin, 2006). In addition, the challenge with qualitative self-reports of emotion is that they are as open to social desirability effects as traditional quantitative self-reports. All self-report measures of emotions require awareness and expression of the emotions experienced and thus severely limit the spectrum of what can be measured and controlled. Observational and physiological measures may help overcome this limitation.

**Observational and Physiological Approaches** Observational approaches do not require participants to be aware of, or report, their emotions. Instead, an observer watches for specific signals, most often including voice (speech), facial expressions, and other intentional behaviors such as tapping of the foot. For example, Hsu et al. (2014) and Leighton (2013) used speech signals to infer distraction and emotional arousal, respectively. However, most of the current work in this area focuses on facial expressions (Reisenzen, Junge, Studtmann, & Huber, 2014). Researchers have developed several systematic rules to infer emotions from the face; for example, the *Emotion Facial Action Coding System* (EMFACS; Ekman, 1972, 1992) and the *Facial Expression Coding System* (FACES; Kring & Sloan, 2007). Although a detailed review of these rules is beyond the scope of the present chapter, their application has been found to lead to high inter-rater reliability (Gottman & Levenson, 2002; Kring & Sloan, 2007). However, their validity to detect many emotions with the exception of amusement tends to be moderate to low (Kring & Sloan, 2007; Reisenzen, Junge, Studtmann, & Huber, 2014). Aside from EMFACS and FACES rules, intuitive inferences of emotions from facial expressions can also be used. However, intuitive inferences tend to lead to lower reliability and validity estimates (Reisenzen et al., 2014). Reliability can be increased by pooling the judgments of several observers (Rosenthal, 2005).

Observational approaches are time-consuming, and thus, in practice, are often only applied to small numbers of interview participants. One exciting new research area is the emergence of computer-based observational programs, which have the

potential to increase reliability and greatly reduce analysis time. Currently, there are two real-time automatic facial coding systems available: FaceReader™ (D’Arcey, Johnson, & Ennis, 2012) and FACET™ (Littlewort et al., 2011). Also, computer programs are being developed to infer emotions from voice patterns (Zeng, Pantic, Roisman, & Huang, 2009), posture (D’Mello & Graesser, 2009) and physiological reactions (Calvo & D’Mello, 2010). While these programs may not currently perform as well as human judgments, computer programs are expected to eventually overtake human observational approaches in the near future (Reisenzen et al., 2014).

In addition to observational approaches, physiological or biometric measures of brain, heart rate and skin conductance could be used to evaluate emotional changes during think-aloud interviews. Although older biometric tools were large and therefore intrusive and awkward for use, recent tools provide extraordinary agility. For example, recent tools such as NeuroSky’s Brainwave Sensing headset to measure attention and concentration via EEG activity is worn by participants during interviews as any headset designed for listening to computer instructions or music. Likewise, Empatica’s E4 Wristband provides real-time measurements of motion-based activity, electro-dermal activity (i.e., arousal of the sympathetic nervous system to derive features related to stress, engagement and excitement), peripheral skin temperature, and blood volume pulse, from which to evaluate heart rate, heart rate variability, and other cardiovascular changes. Again, the wristband can be worn easily by participants during an interview for unobtrusive measurements. To be sure, these tools are expensive as they are commercial products with significant research going into their development. However, less expensive but experimental developments include smart phone apps that also allow for non-intrusive physiological measures. For example, currently, one of the most popular apps for monitoring heart rate is Azumio Inc.’s *Instant Heart Rate* app for both android phone and iPhone. While the accuracy of the apps need to be further evaluated, they have outstanding potential for measuring physiological responses in think-aloud interview research.

## **Integrating Evidence for Validity Arguments: Accounting for Emotions in Think-Aloud Interviews**

The integration of response processing data into validity arguments to defend test-based inferences about test-takers’ cognitive processing is an opportunity for rigor but also a potential trap. On the one hand, it is an opportunity for rigor because it permits test developers to collect relevant evidence – response processing data – that one assumes is related to the construct – test-takers’ cognitive processes – in order to bolster inferences about whether test items do indeed elicit the processing expected in correct and incorrect answers. Of course, the latter assumes the test was designed to measure some aspect of cognitive processing. On the other hand, it is a potential a trap because the gathering of low-quality response processing data may provide a semblance of due course and diligence but, in fact, reveal little of

substance in relation to the validation of inferences and possibly even misinform efforts. Low-quality response processing data refer to evidence that reflects biased cognitive processing of items by test-takers because of interview conditions that may not actually generalize to normal test-taking circumstances.

### ***Revisiting the Value-Added of Response Processing Data***

At the risk of repeating what has already been written in other publications (e.g., Gorin, 2006; Leighton, 2004), the rationale for collecting response processing data is to provide evidence that test items designed to measure certain, normally higher-level constructs, involving cognitive processes such as scientific or mathematical reasoning are indeed eliciting those processes in test-takers. Other sources of validity evidence include information about test content, internal structure, relations to other variables (convergent and discriminant evidence), criterion, generalization and consequences. All of these sources are more or less important to generating validity arguments, depending on the objectives of the test. However, none of these other sources provide evidence directly about test-takers' real-time, cognitive problem-solving approaches to the items. Thus, response processing data are uniquely able to bolster claims in support of, or in opposition to, claims that items are measuring expected cognitive processing in examinees. However, for the data to be useful, the data must reflect as much as possible the test-taker's response processing to the item and not the conditions in which the data were gathered.

Consider again the results of Leighton's (2013) study. High to moderately-high ability students who were led to believe they were participating in think-aloud interviews conducted by an *expert* or who were not told anything about the expertise of the interviewer (i.e., control students) were observed to have statistically *lower* scores on the items than those who thought they were being interviewed by a novice. Although students were randomly assigned to conditions and there were no differences among students in self-reported anxiety, self-monitoring/regulation, confidence and familiarity with content, these two think-aloud conditions (i.e., expert and control) *depressed* the response processing of students – most likely from the activation of some type of unconscious negative emotion (see Berridge & Winkelman, 2003). Incorrect responses can be inferred to originate from lack of knowledge or slips in cognitive processing. Interestingly, compared to the students in the novice condition, the expert/control condition students were found to present cognitive models of equal skill and sophistication in concurrent verbal reports but the execution of processing was not done properly. In retrospective reports, they presented cognitive models of significantly *lower* skill and sophistication compared to students in the novice condition. If the data from the expert/control students were being used to validate items, it might lead to the conclusion that the items were not functioning properly.

### ***Controlling and Minimizing Effects of Emotion: Avoiding the Trap***

There is value in recognizing the potential impact of emotional and other types of confounding variables in the gathering of response processing data in so far as efforts can be undertaken to improve procedures, quality of data and thus evidence for validity. Research in this area of validity is needed. However, several trial strategies can be considered for use in the planning and execution of think-aloud studies to improve the quality of data. These strategies include not only aiming to control for emotions by measuring their prevalence (see earlier section **Approaches for Accounting and Controlling for Emotion**) but also minimizing the effects of negative emotions in the collection of response processing data.

Although studies of emotion-regulation in educational contexts abound, we are not aware of any studies that have directly examined emotion regulation in think-aloud interviews. Thus, much of the foregoing are potential strategies for minimizing the effects of emotion in disrupting cognitive processing in response processing data that require further empirical investigation. First, one way to minimize the biasing effects of emotion in think-aloud studies may be to consider having *computer-generated facilitators* of the think-aloud interview. By removing the presence of a human interviewer, who is perceived not only as an observer but possibly also as a judge, participants' negative emotional responses may be lessened and calibrated back to more traditional test-taking situations. Second, another way to minimize disruptive emotional responses may be to help participants self-regulate their emotions in this new context in much the same way they would in a normal test-taking situation.

Within educational contexts, Jacobs and Gross (2014) classified emotion regulation techniques into five categories based on their *Modal Model of Emotion*; namely, situation selection, situation modification, attentional deployment, cognitive change, and response modulation. For example, situational modification techniques might include positioning the interviewer and participant side-by-side rather than face-to-face, as the latter can cause more anxiety for an interviewee than side-by-side or right angle seating (Osato & Ogawa, 2003). A caution implementing any technique, however, is to consider whether the technique could in fact interfere with the objective of the interview. For example, attention deployment techniques, which divert a person's attention to certain aspects of a situation to reduce their unease (see Gross & Thompson, 2007), could backfire if the student's attention is diverted away from the main objective of the think-aloud interview, such as solving the task of interest.

Other techniques that hold promise involve cognitive change and response modulation. Cognitive change refers to "changing how one appraises the situation one is in so as to alter its emotional significance, either by changing how one thinks about the situation or about one's capacity to manage the demands it poses" (Gross & Thompson, 2007, p. 20). For example, in a recent study, Chu and Leighton (2016) reported that instructors, who explained to undergraduate students the value of

making mistakes during learning, found their students reported (a) a higher level of wellbeing in the classroom, (b) a higher number of reported errors during skill acquisition, and (c) a greater tendency to correct their mistakes using peer feedback. In addition, investigators (e.g., Ben-Zeev, Fein, & Inzlicht, 2005; Jamieson, Mendes, Blackstock, & Schmader, 2010; Johns, Inzlicht, & Schmader, 2008) have looked at helping students to reframe anxiety as beneficial to their performance since mild anxiety may actually boost performance by making cognition more detail oriented; students who adopt this view have been found to improve their test performance. Response modulation is defined by Gross and Thompson (2007, p. 22) as “influencing physiological, experiential, or behavioral responding as directly as possible.” For example, Carney, Cuddy, and Yap (2010) showed that adopting an open physical posture during activities can increase testosterone, decrease cortisol, and lead to subsequent feelings of power and tolerance for risk, while a close posture can result in the opposite. Cuddy, Wilmuth, Yap, and Carney (2015) showed that open postures can boost participants’ performance in a job interview situation. Research on how the instructions to think-aloud interviews could be elaborated and manipulated with inclusion of such variables to relieve anxiety has not been done and should be pursued.

## Conclusions

Given the increasing prevalence of employing think aloud interviews to collect response processing data, it is imperative for researchers to consider the variables that may impact the quality of data generated. Although the Standards (AERA, APA, & NCME, 2014) indicate the benefit of collecting response processing data to bolster validity arguments, very little is still known about the quality of these data, the methodological procedures and interview conditions for generating the highest quality data possible, and even best practices for coding, interpreting and integrating these data within developing validity arguments. The latter was not the focus of this chapter but nonetheless should be mentioned and considered.

With a few recent exceptions, such as the Achievement Emotions Questionnaire (Pekrun et al., 2002) and Epistemic Emotion Scales (Pekrun & Meier, 2011), measures of academic emotions other than anxiety are still lacking (for a review see Pekrun, & Bühner, 2014), and have not been included in response process studies involving think-aloud interviews. It is notable that in both Leighton (2013) and Hsu et al. (2014), clear expressions of the lack of research in this area are made evident. For example, Hsu et al. (2014, p. 1) state “although the detrimental effects of distraction on a variety of more basic cognitive tasks (e.g., visuo-spatial and working memory tasks) are well known (e.g., Lavie, 2005; Tremblay et al., 2005), studies explicitly examining the effects of distraction on engagement in cognitive-affective think-aloud paradigms are notably absent.” Educational measurement specialists are in a unique and advantageous position to carry out this research given the expertise and rigor we impose on the measurement and interpretation of student



achievement and learning. Validity arguments require not just the collection of data but *high-integrity* data; thus, every effort should be made to scrutinize the quality of the data gathered, including those generated from think aloud interviews.

**Acknowledgement** Preparation of this chapter was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada (SSHRC Grant No. 410-2011-0811). Grantees undertaking such projects are encouraged to express freely their professional judgment. This paper, therefore, does not necessarily represent the positions or the policies of the Canadian government, and no official endorsement should be inferred.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Beilock, S. L., & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General*, *130*, 701–725. doi:10.1037/0096-3445.130.4.701.
- Beilock, S. L., Kulp, C. A., Holt, L. E., & Carr, T. H. (2004). More on the fragility of performance: Choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General*, *133*, 584–600. doi:10.1037/0096-3445.133.4.584.
- Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology*, *41*, 174–181. doi:10.1016/j.jesp.2003.11.007.
- Berridge, K. C., & Winkielman, P. (2003). What is an unconscious emotion? (The case for unconscious “liking”). *Cognition and Emotion*, *17*, 181–211. doi:10.1080/02699930244000273.
- Beukeboom, C. J., & Semin, G. R. (2006). How mood turns on language. *Journal of Experimental Social Psychology*, *42*, 553–566. doi:10.1016/j.jesp.2005.09.005.
- Bishop, S. J. (2007). Neurocognitive mechanisms of anxiety: An integrative account. *Trends in Cognitive Sciences*, *11*, 307–316. doi:10.1016/j.tics.2007.05.008.
- Bosson, J. K., Haymovitz, E. L., & Pintel, E. C. (2004). When saying and doing diverge: The effects of stereotype threat on self-reported versus non-verbal anxiety. *Journal of Experimental Social Psychology*, *40*, 247–255. doi:10.1016/S0022-1031(03)00099-4.
- Calvo, R. A., & D’Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37.
- Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*, 1363–1368. doi:10.1177/0956797610383437.
- Carver, C. S., Peterson, L. M., Follansbee, D. J., & Scheier, M. F. (1983). Effects of self-directed attention on performance and persistence among persons high and low in test anxiety. *Cognitive Therapy and Research*, *7*, 333–353. doi:10.1007/BF01177556.
- Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*, 270–295. doi:10.1006/ceps.2001.1094.
- Chu, M.-W., Guo, Q., & Leighton, J. P. (2014). Students’ interpersonal trust and attitudes towards standardized tests: Exploring affective variables related to student assessment. *Assessment in Education: Principles, Policy & Practice*, *21*, 167–192. doi:10.1080/0969594X.2013.844094.
- Chu, M.-W., & Leighton, J. P. (2016). Using errors to enhance learning feedback in computer programming. In S. Tettegah & M. McCreery (Eds.), *Emotions and technology: Communication of feelings for, with and through digital media – Volume I: Emotions, learning, and technology* (pp. 89–117). Elsevier Publishing.



- Cizek, G. J., & Burg, S. S. (2006). *Addressing test anxiety in a high-stakes environment: Strategies for classrooms and schools*. Thousand Oaks, CA: Corwin Press.
- Cohen, G. L., & Sherman, D. K. (2005). Stereotype threat and the social and scientific contexts of the race achievement gap. *American Psychologist*, *60*, 270–271. doi:[10.1037/0003-066X.60.3.270](https://doi.org/10.1037/0003-066X.60.3.270).
- Cuddy, A. J., Wilmuth, C. A., Yap, A. J., & Carney, D. R. (2015). Preparatory power posing affects nonverbal presence and job interview performance. *Journal of Applied Psychology*, *100*, 1286–1295. doi:[10.1037/a0038543](https://doi.org/10.1037/a0038543).
- D'Arcey, T., Johnson, M., & Ennis, M. (2012). Assessing the validity of FaceReader using facial electromyography. In *Proceedings of APS 24th annual meeting*. Retrieved from [www.darcey.us/pdf/facereader.pdf](http://www.darcey.us/pdf/facereader.pdf)
- D'Mello, S., & Graesser, A. (2009). Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence*, *23*, 123–150. doi:[10.1080/08839510802631745](https://doi.org/10.1080/08839510802631745).
- DeBellis, V. A., & Goldin, G. A. (2006). Affect and meta-affect in mathematical problem solving: A representational perspective. *Educational Studies in Mathematics*, *63*, 131–147. doi:[10.1007/s10649-006-9026-4](https://doi.org/10.1007/s10649-006-9026-4).
- Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation* (Vol. 19, pp. 207–283). Lincoln, NE: University of Nebraska Press.
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*, 169–200. doi:[10.1080/02699939208411068](https://doi.org/10.1080/02699939208411068).
- Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 223–241). Cambridge, UK: Cambridge University Press. doi:[10.2277/0521600812](https://doi.org/10.2277/0521600812).
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.
- Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*, 336–353. doi:[10.1037/1528-3542.7.2.336](https://doi.org/10.1037/1528-3542.7.2.336).
- Fiedler, K. (2001). Affective states trigger processes of assimilation and accommodation. In L. L. Martin & G. L. Clore (Eds.), *Theories of mood and cognition: A user's guidebook* (pp. 85–98). Mahwah, NJ: Lawrence Erlbaum.
- Fiedler, K., & Beier, S. (2014). Affect and cognitive processes in educational contexts. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 348–367). New York, NY: Yarrow & Francis.
- Forgas, J. P. (1998). On feeling good and getting your way: Mood effects on negotiator cognition and bargaining strategies. *Journal of Personality and Social Psychology*, *74*, 565–577. doi:[10.1037//0022-3514.74.3.565](https://doi.org/10.1037//0022-3514.74.3.565).
- Frenzel, A. C. (2014). Teacher emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 494–519). New York, NY: Taylor & Francis.
- Goldin, G. A. (2000). Affective pathways and representation in mathematical problem solving. *Mathematical Thinking and Learning*, *2*, 209–219. doi:[10.1207/S15327833MTL0203\\_3](https://doi.org/10.1207/S15327833MTL0203_3).
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*, 21–35. doi:[10.1111/j.1745-3992.2006.00076.x](https://doi.org/10.1111/j.1745-3992.2006.00076.x).
- Gottman, J., & Levenson, R. W. (2002). A two-factor model for predicting when a couple will divorce: Exploratory analyses using 14-year longitudinal data. *Family Process*, *41*, 83–96. doi:[10.1111/j.1545-5300.2002.40102000083.x](https://doi.org/10.1111/j.1545-5300.2002.40102000083.x).
- Gross, J. J., & Thompson, R. A. (2007). Emotion regulation: Conceptual foundations. In J. J. Gross (Ed.), *Handbook of emotion regulation* (pp. 3–24). New York, NY: Guildford Press.
- Hoehn-Saric, R., & McLeod, D. R. (2000). Anxiety and arousal: Physiological changes and their perception. *Journal of Affective Disorders*, *61*, 217–224. doi:[10.1016/S0165-0327\(00\)00339-6](https://doi.org/10.1016/S0165-0327(00)00339-6).

- Hsu, K. J., Babeva, K. N., Feng, M. C., Hummer, J. F., & Davison, G. C. (2014). Experimentally induced distraction impacts cognitive but not emotional processes in think-aloud cognitive assessment. *Frontiers in Psychology*, *5*, 1–9. doi:[10.3389/fpsyg.2014.00474](https://doi.org/10.3389/fpsyg.2014.00474).
- Huntsinger, J. R., Clore, G. L., & Bar-Anan, Y. (2010). Mood and global–local focus: Priming a local focus reverses the link between mood and global–local processing. *Emotion*, *10*, 722–726. doi:[10.1037/a0019356](https://doi.org/10.1037/a0019356).
- Jacobs, S. E., & Gross, J. J. (2014). Emotion regulation in education: Conceptual foundations, current applications, and future directions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 348–367). New York, NY: Yalor & Francis.
- Jamieson, J. P., Mendes, W. B., Blackstock, E., & Schmader, T. (2010). Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE. *Journal of Experimental Social Psychology*, *46*, 208–212. doi:[10.1016/j.jesp.2009.08.015](https://doi.org/10.1016/j.jesp.2009.08.015).
- Johns, M., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General*, *137*, 691–705. doi:[10.1037/a0013834](https://doi.org/10.1037/a0013834).
- King, G., & Wand, J. (2007). Comparing incomparable survey responses: New tools for anchoring vignettes. *Political Analysis*, *15*, 46–66. doi:[10.1093/pan/mpj011](https://doi.org/10.1093/pan/mpj011).
- Kleinginna Jr., P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, *5*, 345–379. doi:[10.1007/BF00992553](https://doi.org/10.1007/BF00992553).
- Kring, A. M., & Sloan, D. M. (2007). The facial expression coding system (FACES): Development, validation and utility. *Psychological Assessment*, *19*, 210–224. doi:[10.1037/1040-3590.19.2.210](https://doi.org/10.1037/1040-3590.19.2.210).
- Kyllonen, P. (2016). Socio-emotional and self-management variables in learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), *Handbook of cognition and assessment* (pp. 174–197). Wiley-Blackwell.
- Kyllonen, P. C., & Bertling, J. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–285). Boca Raton, FL: CRC Press. doi:[10.1111/jedm.12095](https://doi.org/10.1111/jedm.12095).
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*, 6–15. doi:[10.1111/j.1745-3992.2004.tb00164.x](https://doi.org/10.1111/j.1745-3992.2004.tb00164.x).
- Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education*, *26*, 136–157. doi:[10.1080/08957347.2013.765435](https://doi.org/10.1080/08957347.2013.765435).
- Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education. Theory and applications*. Cambridge, MA: Cambridge University Press. doi:[10.1111/j.1745-3984.2008.00072.x](https://doi.org/10.1111/j.1745-3984.2008.00072.x).
- Lewis, B., & Linder, D. (1997). Thinking about choking? Attentional processes and paradoxical performance. *Personality and Social Psychology Bulletin*, *23*, 937–944. doi:[10.1177/0146167297239003](https://doi.org/10.1177/0146167297239003).
- Littlewort, G., Whitehill, J., Wu, T., Fasel, I. R., Frank, M., Movellan, J. R., & Bartlett, M. S. (2011). The computer expression recognition toolbox (CERT). In *Proceedings of the 9th IEEE conference on Automatic Face and Gesture Recognition* (pp. 298–305), Santa Barbara, CA.
- Maehr, M. L., & Meyer, H. A. (1997). Understanding motivation and schooling: Where we've been, where we are, and where we need to go. *Educational Psychology Review*, *9*, 371–409. doi:[10.1023/A:1024750807365](https://doi.org/10.1023/A:1024750807365).
- Malmivuori, M. L. (2006). Affect and self-regulation. *Educational Studies in Mathematics*, *63*, 149–164. doi:[10.1007/s10649-006-9022-8](https://doi.org/10.1007/s10649-006-9022-8).
- Mayer, J. D., & Gaschke, Y. N. (1988). The experience and meta-experience of mood. *Journal of Personality and Social Psychology*, *55*, 102–111. doi:[10.1037/0022-3514.55.1.102](https://doi.org/10.1037/0022-3514.55.1.102).

- McCulloch, A. W. (2011). Affect and graphing calculator use. *The Journal of Mathematical Behavior*, *30*, 166–179. doi:10.1016/j.jmathb.2011.02.002.
- Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*, 119–124. doi:10.1177/1754073912468165.
- Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, *64*, 723–739. doi:10.1037//0022-3514.64.5.723.
- Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking performance. *Journal of Educational Measurement*, *27*, 41–58. doi:10.1111/j.1745-3984.1990.tb00733.x.
- Osato, E., & Ogawa, N. (2003). Effects of seating positions on heart rates, state anxiety, and estimated interview duration in interview situations. *Psychological Reports*, *93*, 755–770. doi:10.2466/pr0.2003.93.3.755.
- Paulhus, D. L. (1991). Measurement and control of response biases. In J. Robinson, P. Shaver, & L. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17–59). San Diego, CA: Academic Press.
- Pekrun, R. (1992). The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology*, *41*, 359–376. doi:10.1111/j.1464-0597.1992.tb00712.x.
- Pekrun, R., & Bühner, M. (2014). Self-report measures of academic emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 348–367). New York, NY: Yarlor & Francis.
- Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control-value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, *102*, 531–549. doi:10.1037/a0019243.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist*, *37*, 91–106. doi:10.1207/S15326985EP3702\_4.
- Pekrun, R., & Linnenbrink-Garcia, L. (Eds.). (2014). *International handbook of emotions in education*. New York, NY: Routledge.
- Pekrun, R., & Meier, E. (2011). *Epistemic Emotion Scales (EES)*. Unpublished manuscript, Department of Psychology, University of Munich, Munich, Germany.
- Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. (2007). *The development and psychometric properties of LIWC 2007*. Retrieved from <http://www.liwc.net/LIWC2007LanguageManual.pdf>
- Piaget, J. (1954). Language and thought from a genetic perspective. *Acta Psychologica*, *10*, 51–60. doi:10.1016/0001-6918(54)90004-9.
- Reisenzein, R., Junge, M., Studtmann, M., & Huber, O. (2014). Observational approaches to the measurement of emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 348–367). New York, NY: Yarlor & Francis.
- Rosenthal, R. (2005). Conducting judgment studies: Some methodological issues. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *New handbook of methods in nonverbal behavior research* (pp. 199–234). New York, NY: Oxford University Press.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*, 145–172. doi:10.1037/0033-295X.110.1.145.
- Ryan, K. E., & Ryan, A. M. (2005). Psychological processes underlying stereotype threat and standardized math test performance. *Educational Psychologist*, *40*, 53–63. doi:10.1207/s15326985ep4001\_4.
- Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks*, *18*, 317–352. doi:10.1016/j.neunet.2005.03.001.
- Sarason, I. G., Sarason, B. R., & Pierce, G. R. (1995). Cognitive interference: At the intelligence-personality crossroads. In D. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 285–296). New York, NY: Plenum Press.

- Sawyer Jr., T. P., & Hollis-Sawyer, L. A. (2005). Predicting stereotype threat, test anxiety, and cognitive ability test performance: An examination of three models. *International Journal of Testing*, *5*, 225–246. doi:[10.1207/s15327574ijt0503\\_3](https://doi.org/10.1207/s15327574ijt0503_3).
- Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. New York, NY: Oxford University Press.
- Shuman, V., & Scherer, K. R. (2014). Concepts and structures of emotions. In *International handbook of emotions in education* (pp. 13–35). New York, NY: Routledge. doi:[10.4324/9780203148211.ch2](https://doi.org/10.4324/9780203148211.ch2).
- Sinclair, R. C., & Mark, M. M. (1995). The effects of mood state on judgmental accuracy: Processing strategy as a mechanism. *Cognition & Emotion*, *9*, 417–438. doi:[10.1080/02699939508408974](https://doi.org/10.1080/02699939508408974).
- Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, E. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Manual for the test anxiety inventory* (“Test Attitude Inventory”). Redwood City, CA: Consulting Psychologists Press.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629. doi:[10.1037//0003-066X.52.6.613](https://doi.org/10.1037//0003-066X.52.6.613).
- Storbeck, J., & Clore, G. L. (2005). With sadness comes accuracy; With happiness, false memory mood and the false memory effect. *Psychological Science*, *16*, 785–791. doi:[10.1111/j.1467-9280.2005.01615.x](https://doi.org/10.1111/j.1467-9280.2005.01615.x).
- Storbeck, J., & Clore, G. L. (2008). The affective regulation of cognitive priming. *Emotion*, *8*, 208–215. doi:[10.1037/1528-3542.8.2.208](https://doi.org/10.1037/1528-3542.8.2.208).
- Turner, J. C., Christensen, A., Kackar-Cam, H. Z., Trucano, M., & Fulmer, S. M. (2014). Enhancing students’ engagement: Report of a 3-Year intervention with middle school teachers. *American Educational Research Journal*, *51*, 1195–1226. doi:[10.3102/0002831214532515](https://doi.org/10.3102/0002831214532515).
- Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*, 459–482. doi:[10.1002/cne.920180503](https://doi.org/10.1002/cne.920180503).
- Zeidner, M. (2014). Anxiety in education. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 265–288). New York City, NY: Routledge.
- Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*, 39–58. doi:[10.1109/TPAMI.2008.52](https://doi.org/10.1109/TPAMI.2008.52).
- Zuriff, G. E. (1997). Accommodations for test anxiety under ADA. *The Journal of the American Academy of Psychiatry and the Law*, *25*, 197–206.

# Chapter 9

## Response Time Data as Validity Evidence: Has It Lived Up To Its Promise and, If Not, What Would It Take to Do So

Zhi Li, Jayanti Banerjee, and Bruno D. Zumbo

### What Is Response Time and How Can It Be Used for Validity Evidence?

The widespread use of computers in test delivery provides easy access to data on response processes; it is possible to track and trace events like clicks of the mouse, movement of text or objects, or the time that test takers use to respond to items (i.e., response time). Response time (RT) is typically defined as the time a test taker uses to complete an item or task, beginning from the initial presentation of the task to the time at which the complete response is logged. RT has attracted substantial attention in recent years as it offers a promising window into test takers' cognitive processes and thus the construct being measured (Huff & Sireci, 2001; Schnipke & Scrams, 1997). Molenaar (2015) argues that "the natural variability in response times can give valuable information for psychological and educational inferences about response processes and solution strategies" (p. 177). As such, RT offers an opportunity to build validity evidence for a test.

This expectation that RT data can and should be a source of validity evidence originates in part from its association with response processes; the time that a test taker takes to process and respond to an item is a natural corollary of the psychological, cognitive or thinking processes activated during the act of responding to an item. Indeed, the *Standards for Educational and Psychological Testing* (AERA,

---

Z. Li (✉) • J. Banerjee

Paragon Testing Enterprises, Inc., 110-2925 Virtual Way, Vancouver, BC V5M 4X5, Canada  
e-mail: [zli@paragontesting.ca](mailto:zli@paragontesting.ca); [jbanerjee@paragontesting.ca](mailto:jbanerjee@paragontesting.ca)

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: [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca)

© Springer International Publishing AG 2017

B.D. Zumbo, A.M. Hubleby (eds.), *Understanding and Investigating Response Processes in Validation Research*, Social Indicators Research Series 69,  
DOI 10.1007/978-3-319-56129-5\_9

APA, & NCME, 2014) lists response processes as one of five sources of validity evidence, others being content-related, internal structure, associations with other variables, and consequences. Despite this, response processes have not yet been adequately investigated in the field of educational measurement. Lyons-Thomas, Liu, and Zumbo (2014) comment that the use of response processes as validity evidence has been “virtually ignored” (p. 316). Additionally, out of a collection of 15 chapters which offer syntheses of validity and validation activity in the social, behavioral, and health sciences (Zumbo & Chan, 2014) only eight cite studies on response processes. Moreover, the number of such studies is painfully small; of the studies included, the percentage focusing on response processes did not top 9.5%. Therefore, even though the number of studies on RT data is on the rise, more studies, especially those analyzing response data for validity evidence, are still needed.

In the sections that follow, we briefly present background information about RT in assessment contexts and the application of RT data to solving real-life issues. We then review empirical studies of RT both from modeling and non-modeling perspectives. We conclude the chapter with an evaluation of RT data as a source of validity evidence, addressing the question of whether RT has yet lived up to its potential. Since we believe that more can be achieved, we offer suggestions for how RT research might live up to its promise.

## **What Do Response Time Data in Assessment Contexts Tell Us?**

In cognitive psychology, RT is also known as reaction time (Ratcliff, van Zandt, & McKoon, 1999) and response latency (Mislevy, 1989; Parshall, Mittelholtz, & Miller, 1994; Ranger & Ortner, 2012; Siem, 1996). These terms have subtle distinctions that are not essential for our purposes so we will consider them interchangeable and use the term ‘response time’ throughout.

RT is not a new data type in research. The nineteenth century German scientist Hermann von Helmholtz may be one of the earliest pioneers of psychological research using RT in his experiments with the nerve impulse in frog legs. In 1938, the use of RT was discussed and practiced by Woodworth in a series of seminal studies in cognitive psychology, which laid the foundation for further employment of RT in perceptual-motor tasks starting in the 1960s (Ratcliff et al., 1999).

In the field of educational measurement, RT refers to the time a test taker or respondent spends answering an item and it can be conveniently operationalized as the temporal interval captured from the timestamp of stimulus or item presentation to the timestamp when a test taker registers his or her response to the item. The attention to RT in measurement can be traced back to the 1950s, as witnessed by Gulliksen’s (1950) distinction between speed and power tests. A pure speed test would consist of easy items whose probability of being correctly answered under untimed conditions is almost 1 (100%). However, such a test would require test takers to respond within a limited span of time. Therefore, the test takers’ level of the



trait of interest is reflected in their speed of responding to the items. By contrast, a pure power or ability test would consist of items covering a range of difficulty levels and test takers would be given unlimited time to respond. In such a power test, the test takers' trait level would be evaluated by the number of items they answered correctly. As Lu and Sireci (2007) observed, most of the educational achievement tests used today do not explicitly describe RT as a part of the construct measured. Therefore, these tests should be viewed as representative of power tests but with a pre-determined generous time limit. They are also called time-limited tests in van der Linden and Hambleton (1997).

Related to the distinction between speed tests and power tests is a common belief about the relationship between response speed and response accuracy in psychological experiments, as well as in (speeded) educational tests. This is known as the speed-accuracy trade-off, meaning that a compromise is made by respondents or test takers between response speed and response accuracy (Dennis & Evans, 1996; Dodonova & Dodonova, 2013; Wickelgren, 1977). Concerns over the impact of the speed-accuracy trade-off on the meaning of test scores have prompted researchers to use RT or speed information to capture the relationship between response speed and response accuracy (van der Linden, 2009). For this reason, test speededness, broadly defined as the degree to which test performance is affected by the time limit (Schnipke & Scrams, 1997), is an important concern as it introduces construct-irrelevant variance to the tests and poses a threat to the validity of the test score interpretation (Lu & Sireci, 2007).

RT data have been utilized in various stages of test development. For example, Lu and Sireci (2007) list some benefits brought about by using RT in computer-based tests: to better describe response patterns, to improve the precision of parameter estimation, and to identify salient individual traits such as motivation or effort. In the following subsection, we present three situations in which RT analysis plays an important role.

### *Setting Time Limits*

Setting an appropriate time limit is of importance to both test developers and test takers. From a test developer's perspective, the more time allowed for a particular test, the more the test will cost to administer. From a test taker's perspective, longer tests tend to cause fatigue and to increase testing anxiety. In cases where a testing accommodation has been requested (such as for test takers with a visual or hearing impairment), a common strategy has been to provide extra time. The aim of the additional time is to ensure that test takers who have special needs are still given a fair chance of demonstrating their true ability in the trait being evaluated. In theory, an understanding of RT information could help control test speededness for test takers of different abilities or needs, which can ultimately promote test fairness as well as strengthen the interpretation of the scores generated from such tests (Fan, Wang, Chang, & Douglas, 2012; van der Linden, Scrams, & Schnipke, 1999).

Traditional approaches to setting time limits are fairly intuitive or stopwatch-based. A commonly used approach labels a test as unspeeeded if it meets two criteria: at least 80% of the test takers can complete all items and all test takers can complete at least 75% of the items (van der Linden, 2011). Van der Linden operationalized the concept of speededness as an interaction between test takers' speed, the amount of labor needed to respond to an item, and the time limit on a given test. To lend more empirical support to time-constraint setting, he proposed the log-normal model of RT as a new approach to determining in advance (based on pre-calibrated item parameters) the probability of test takers running out of time. Because the distribution of raw RT data tends to be highly skewed, a variety of transformation methods have been applied to RT data and log-transformation is one of the commonly used methods, others being exponential, Gamma, and Weibull distribution (Schnipke & Scrams, 2002; van der Linden, 2009). The lognormal model of RT includes a speed parameter for test takers, a time intensity parameter for the item (i.e., the labor required by the item), and a discrimination parameter. This model can be used to estimate the risk of test takers running out of time on a given test and determine time limits to match a pre-determined risk level.

### *Capturing Aberrant Test-Taking Behaviors*

Another practical use of RT is utilizing pre-determined RT thresholds to identify test takers who exhibit different test-taking behaviors, such as rapid-guessing behaviors or random-guessing behaviors as opposed to solution behaviors (Kong, Wise, & Bhola, 2007; Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; Yang, 2007). Rapid-guessing behaviors happen when test takers rush through a test with little attention to the test content or when they show little interest in the test. Indicators of these behaviors are a shorter RT and usually a lower response-accuracy rate. Solution behaviors, on the other hand, reflect test takers' normal effort in the test and the response times are within expected ranges. These distinctions can account for test takers' motivation levels or the effort made during the test, which are common concerns in low-stakes tests (Wise, 2014; Wise & Kong, 2005).

Kong et al. (2007) discuss different methods of fixing an RT threshold. Using the data from the Information Literacy Test (ILT), a computerized multiple-choice test designed to measure college students' abilities to make use of information from resources such as libraries, they compare four methods: (1) a common threshold for all items (3-s), (2) a threshold based on surface features such as item length and requirement of ancillary reading, (3) visual identification of the threshold based on RT distribution, and (4) a two-state mixture model-based method. Using these methods, Kong et al. classified the item responses that were below the threshold values as potential rapid-guessing behaviors. The other responses were classified as solution behaviors. Kong et al. found that the four methods yielded similar classification results but concluded that the two-state mixture model-based method was more psychometrically rigorous than the other three methods.



## ***Supporting Test Security Practice***

In addition to being important for the valid interpretation of test scores as well as for more precise item parameter estimations (Kong et al., 2007), RT data are useful in the area of test security (Qian, Staniewska, Reckase, & Woo, 2016; van der Linden & Guo, 2008). It can provide the evidential link between aberrant test-taking behavior and over-exposed or compromised items. For example, Qian et al. investigated response aberrances in two computer-based licensure examinations, one being adaptive and another that was non-adaptive. In particular, they demonstrated how the analysis of RT information could identify item pre-knowledge. In contrast to other aberrant test-taking behavior where test takers answer items very quickly but with low accuracy, test takers with item pre-knowledge answer items correctly but with a much shorter RT. Qian et al. followed van der Linden's (2006) approach to modeling RTs using an R package *cirt* (Fox, Klein Entink, & van der Linden, 2007) and modelled item responses with a 2-parameter logistic model for the non-adaptive examination and the Rasch model for the adaptive examination. By comparing the drifts in RT parameters from an early sample and a late sample on the same examination, Qian et al. were able to identify the test takers who might have taken advantage of item pre-knowledge and the items that might have been compromised.

The aforementioned applications of RT indicate the tremendous potential of RT in measurement practice (Kahraman, Cuddy, & Clauser, 2013; Lee & Haberman, 2016; Thomas, 2006). The next section will review a selection of empirical studies using RT with both modeling and non-modeling approaches.

## **How Are Response Time Data Analyzed?**

There are two main approaches to a better understanding of the relationship between test takers' item-level RT and their test performance. One approach is to expand existing models or to create a new model that accommodates RT information when estimating test taker's ability parameters (van der Linden, Klein Entink, & Fox, 2010; Wang, 2005; Wise & DeMars, 2006). Another approach is non-modeling-based and typically uses inferential statistics to explore the relationship between RT and other test-taking related variables (Hess, Johnston, & Lipner, 2013; Lee & Chen, 2011; Lee & Haberman, 2016).

### ***Modeling-Based Approach***

In this subsection, we separate our descriptions of models which are purely statistical and those with substantive cognitive theories or notions. For example, RT information has been incorporated as an additional latent variable in psychometric

models in order to achieve better precision in the measurement of a latent ability or theta (Fox et al., 2007). These models are statistical in nature in that they use latent variables to account for the differences in responses and RT. Alternatively, RT information is combined with response data in cognitive process models such as the Q-diffusion model (van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011), the race model (Ranger, Kuhn, & Garviria, 2015), and the proportional hazards model (Ranger & Kuhn, 2014) to make inferences about response processes. A summary of some major RT models is presented in Table 9.1.

**Table 9.1** Summary of example models with RT components

Category	Model	Main features	Estimation methods and evaluation of model fit	Examples of tests applied
Statistical models	Hierarchical model of responses and RTs (van der Linden, 2007)	The responses and RTs are modelled separately at the first level but simultaneously. The second level features a multivariate normal distribution of test taker ability and speed parameters with multiple correlations among the parameters.	A fully Bayesian approach to parameter estimation with MCMC sampler; DIC, the Bayes factor, and the Bayes residual analysis for model fit evaluation	The computerized CPA Examination
	Bivariate generalized linear item response model (B-GLIRT) (Molenaar, Tuerlinckx, & van der Maas, 2015)	Responses and response times are modeled separately and then linked using a cross-relation function (linear, interaction, and curvilinear) for the latent ability variable and the latent speed variable.	RMSEA, CFI, and TLI for model fit evaluation when WLS estimation is used; AIC, BIC, and sBIC for model fit evaluation when MML estimation is used	Three subscales in the Amsterdam Chess Test (the ACT)
	Mixture Rasch model with RT (MRM-RT) (Meyer, 2010)	This mixture Rasch model uses both item responses and item RT to classify test takers into a solution behavior group and a rapid-guessing behavior group. Class-specific parameters are then estimated.	MCMC procedure for Bayesian estimation	The Information Literacy Test (ILT)

(continued)

**Table 9.1** (continued)

Category	Model	Main features	Estimation methods and evaluation of model fit	Examples of tests applied
Substantive models	Mixture proportional hazards model (Ranger & Kuhn, 2013, 2014, 2016)	Latent class analysis and class-specific proportional hazards models with random effects are used to model RTs. Differences in the cognitive processes are reflected in comparisons with class-specific baseline hazards functions.	MML estimation; chi-squared like test and information criteria (AIC and BIC) for model fit evaluation	A mental calculation test for 2nd-4th grade students
	Race model (Ranger, Kuhn, & Gaviria, 2015)	The model postulates two latent traits, namely general response speed and preference of one of the options, which account for two accumulation processes (information vs. misinformation) competing with each other to reach their thresholds.	Maximum a posteriori (MAP) estimator; Likelihood ratio test and test of item fit for model fit evaluation	Two subscales in the Amsterdam Chess Test (the ACT)
	Diffusion item response model (van der Maas et al., 2011)	RT is a function of drift rate, boundary separation, and non-decision time. Drift rate and boundary separation in traditional diffusion models are decomposed into item-specific and person-specific parameters in the diffusion IRT models.	MML for parameter estimation; absolute and comparative model fit indices, including $M_r$ statistic, QQ plot, AIC, BIC, sBIC, DIC, and likelihood ratio test.	Extraversion data for D-diffusion model; Mental rotation data for Q-diffusion model
	Distance-based IRT model (Ferrando & Lorenzo-Seva, 2007)	Log RT is treated as a function of weighted person-distance measure (based on the distance-difficulty hypothesis in personality measurement), which includes item difficulty, discrimination, and person ability, with additional RT-related parameters.	Iterative procedure and the MML procedure using EM algorithm for parameter estimation; Residual analysis and graphic procedures for model fit evaluation	Binary items in two short scales from personality questionnaires

Note: *MCMC* Markov chain Monte Carlo, *DIC* deviance information criterion, *RMSEA* root mean square error of approximation, *CFI* comparative fit index; *TLI* Tucker-Lewis index, *QQ plot* quantile-quantile plot, *AIC* akaike information criterion; *BIC* Bayesian information criterion, *sBIC* Schwarz' BIC, *MML* marginal maximum likelihood, *EM* The expectation-maximization algorithm, *CPA* certified public accountant

## Statistical Models of Response Time

Lu and Sireci (2007) distinguish two categories of RT models: the ones focusing on RT as a classifying variable and the ones treating RT as a source of information for parameter estimation. Schnipke and Scrams (1997) is one of the earliest studies that focuses on RT as a classifying variable. The study investigated the relationship between RT distributions and test-taking behaviors (solution behavior vs. rapid-guessing behavior) in educational measurement. Schnipke and Scrams observed that a bimodal distribution of RT is indicative of different test-taking behaviors, which is particularly common for the items appearing at the end of a test. They used a lognormal distribution model to accommodate the typical positive skewness in RT data and estimated the scale of RT (median of RT), the shape (standard deviation of RT), and the proportion of rapid-guesses on each item. To evaluate their mixture models, Schnipke and Scrams fitted three RT models to the individual items in the analytical section of a non-adaptive computerized GRE General Test, namely, the single-statement model, two-state mixture model, and common-guessing mixture model based on the distributional characteristics of RT. Their comparative analysis of the root mean squared error (RMSE) suggests that the two-state mixture model showed best fit to the data. Consequently, test speededness can be gauged with the proportion of rapid guesses as well as the proportion of the test takers who did not complete the test. It is noteworthy that Schnipke and Scrams modeled RT only and did not attempt to incorporate RT with item parameter estimation.

Another group of models focusing on the functionality of RT as a classifying variable have adopted a mixture feature which allows for separate parameter estimates for heterogeneous groups. The argument is that if differences in test taker groups are ignored, this tends to cause biases in parameter estimation. Therefore, mixture models employ RT data to classify test takers who may exhibit different test-taking behaviors. An example of this approach is Meyer (2010). Meyer developed a mixture Rasch model with item RT (MRM-RT) to estimate item parameters separately for two latent classes of test takers, namely a rapid-guessing behavior group and a solution behavior group, based on the item RT data. Similar to the RT model in van der Linden's framework, the RT data in Meyer's model are assumed to follow the lognormal distribution, but with latent class indicator information. Meyer applied his model to test data from the Information Literacy Test (ILT). He found that 15% of the test takers of the ILT showed rapid-guessing behavior and the MRM-RT model with two latent classes performed better than a non-mixture model. Meyer also reported the results from a simulation study showing the superiority of the two-class model over the one-class model in terms of accuracy in parameter estimation. According to Meyer, the model can potentially accommodate more latent classes of test takers based on their behavior patterns. However, the Bayesian estimation method used in MRM-RT is extremely time-consuming; the simulation analysis took 2–3 weeks to run. This suggests that the model could not be easily applied in an operational testing program.

More recently, Chan, Lu, and Tsai (2014) introduced a new class of test-taking behavior, namely, higher ability and responding with familiarity (HARF) behavior.

To explore this test-taking behavior, Chan et al. took a mixture structural equation modeling approach to test the same ILT data used in Meyer (2010) with a three-class mixture Rasch model. Using maximum likelihood with robust standard errors (MLR) as the estimation method, Chan et al. found that the three-class model showed good comparative model fit in relation to the 2-class model and the 4-class model in terms of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), as well as good absolute model fit in terms of the relationship between observed and expected probabilities of correct responses from the test takers in the three classes.

Rather than relying on RT data for classification purposes, other psychometric models treat RT as the manifestation of latent variables and use it alongside response data to estimate both item and person parameters. Among the psychometric models of RT, the hierarchical framework for modeling speed and accuracy proposed by van der Linden (2007) is regarded as the “most promising approach within IRT” (van der Maas et al., 2011, p. 351). Van der Linden’s framework features two levels of models – a response model and a RT model at the first level to estimate person and item parameters, and a population or person model and an item-domain model at the second level to represent the relations between the parameters from the level 1 models. Specifically, the person parameters consist of a speed parameter and a person ability parameter, while the item parameters include an item intensity parameter plus the usual IRT parameters, namely, item difficulty, discrimination, and guessing, depending on the types of response models. Van der Linden’s framework takes a “plug-and-play approach”, i.e., different types of models for responses or RTs may be included at the first level and different types of models for parameters at the second level in this framework. For example, in van der Linden’s example application of the hierarchical framework to empirical data, a three-parameter normal-ogive model of response and a lognormal model of RT were chosen at the first level, and multivariate normal models were used at the second level. The R package *cirt* is designed to estimate parameters in a conjoint IRT model of both responses and RTs in a hierarchical model framework (Fox et al., 2007). Nevertheless, it should be noted that this hierarchical framework is premised on some strong assumptions. For example, the assumption of stationarity states that a test taker’s response speed is fixed for all items. Another relevant assumption is the conditional independence between responses and RT given an ability level and a speed level. In most testing situations these might not be safe assumptions to make.

A number of variant models have been developed to jointly model both response and RT under the hierarchical framework (Klein Entink, Kuhn, Hornke, & Fox 2009; Meng, Tao, & Chang, 2015; van der Linden et al., 2010; Wang, Chang, & Douglas, 2013). Scherer, Greiff, and Hautamaki (2015) adopted van der Linden’s (2007) hierarchical framework to investigate the relationship between RT and test taker ability in complex problem solving. The R package *cirt* was used by Scherer et al. to jointly model both item responses and RTs from 2000 Finnish students on nine tasks. The output indicators, such as the expected-a-posteriori (EAP) estimates of the RT and the latent trait of complex problem-solving ability from the joint model, were then fed to a structural equation model to examine the impact of other

individual traits (e.g., goal orientation) on RTs as well as the role of RT in predicting school achievement. As a generalization of van der Linden's (2007) hierarchical model, Meng et al. (2015) proposed a conditional joint model of item responses and RTs. Their aim was to account for the local dependency between speed and accuracy, while observing a relaxed assumption of the conditional independence between item responses and RT. In Meng et al.'s model, the two-level framework is enriched by two correlation structures between item and person parameters so that the new model is flexible with respect to the relationship between speed and accuracy. Meng et al.'s simulation study suggests that the conditional joint model provides less biased and more efficient estimates of the latent trait in cases where the independence assumption is violated.

Recently, Molenaar et al. (2015) proposed a bivariate generalized linear item response model (B-GLIRT) which consists of a measurement model of responses and a measurement model of RTs. To link these two models, Molenaar et al. defined a cross-relation function to reflect the relationship between the latent ability variable and the latent speed variable. Molenaar et al. claim that other IRT-based response models, such as van der Linden's hierarchical model (discussed earlier) and Ferrando and Lorenzo-Seva's (2007) IRT model, can be thought of as special cases of B-GLIRT, through specifying different cross-relations (e.g., linear, interaction, and curvilinear). In this way, these IRT-based models of response and RT can be compared directly.

### **Cognitive Process Models of Response Time**

Cognitive models have only recently been applied in conjunction with response models to reveal the surface-level relationships between RT and simple cognitive processes (Klein Entink et al., 2009). Some cognitive theories about information processing and decision-making have been used as theoretical frameworks to provide explanatory power to RT models, including the diffusion model, the race model, the distance-difficulty hypothesis, and the hazards model.

Diffusion models, a type of sequential sampling model, have been used in decision-making studies in experimental psychology (Ratcliff, 2014; Voss, Nagler, & Lerche, 2013). Diffusion models assume that information accumulates continuously in the process of binary decision-making and such a process can be characterized by the drift rate, which reflects the speed of information accumulation. In the case of binary decision-making, each option requires a certain amount of information to reach a threshold so that a subject would choose that option. The distance between the two boundaries is called boundary separation. Stimulus encoding and response execution are recorded as non-decision time, as they are extra-decisional processes.

Ferrando and Lorenzo-Seva's (2007) IRT model is an expansion of the log-linear RT model proposed by Thissen (1983). They apply the distance-difficulty hypothesis (the DD hypothesis) in the field of personality assessment. The DD hypothesis states that, in the case of a binary decision for a personality trait, the difficulty of

responding to or endorsing the item increases while the person ability-item distance on the same scale narrows. By incorporating a latency submodel for the person-distance measure, Ferrando and Lorenzo-Seva found that their IRT model showed acceptable fit to personality questionnaire data and yielded more accurate parameter estimates and higher test information.

Another example of cognitive process models is the mixture proportional hazards model with random effects proposed by Ranger and Kuhn (2016). The proportional hazards model is a type of survival model and uses hazard functions to investigate event times in medical and biometric research. This model becomes useful when response processes are assumed to be an individual's accumulation of needed information over time whereby the response is made once a threshold of information is met (Ranger & Kuhn, 2015). Therefore, the RT indicates the extent of information accumulation in the response processes and can be used to infer an individual's rate of information acquisition. For example, RT differences found in comparison to the baseline hazards function can be interpreted in light of dual processing theory in cognitive psychology, which distinguishes two modes of information processing, namely a more time-consuming mode of controlled processing and a more efficient mode of automated processing.

Though clearly complex and informative in the context of simulated data or perceptual-motor activities such as chess, it is challenging to apply these cognitive models in educational measurement primarily because of the different natures of the constructs of interest as well as the complexity of tasks used (Klein Entink et al., 2009). A notable exception is Klein Entink et al.'s work. They followed the hierarchical framework and proposed a joint model of response and RT to match item parameters with the underlying design factors or the cognitive skills required to solve each item in a figural matrices test of reasoning ability with non-verbal content. In this model, the cognitive processes or design rules of the items are associated with item difficulty and time intensity, thus providing an approach to evaluating the corresponding cognitive theory in such a rule-based test.

### ***Non-Modeling-Based Approach***

Compared with the modeling-based studies reviewed above, the non-modeling-based studies have primarily focused on the relationship between RT and the characteristics of item content and item formats. Many of the studies in this category have used educational tests as their targets. For example, the variability of RT is usually associated with a number of factors such as content area (Zenisky & Baldwin, 2006), the cognitive complexity of test tasks (Gorin, 2006; Parshall et al., 1994), test taker ability (Hess et al., 2013), and other test taker characteristics such as age and motivation (Wise & Kong, 2005).

Investigating RT in different content areas, Zenisky and Baldwin (2006) studied adult students' item-level RT on a computerized math test and reading test in the U.S. Specifically, Zenisky and Baldwin investigated the relationship between



median RT and a series of test variables and test taker variables, including item difficulty, item complexity, cognitive areas measured by the items, and status of English (as first language or second language). Zenisky and Baldwin showed that the relationships between RT and the characteristics of the test and test takers varied across content areas (reading vs. math) and proficiency levels in the respective content area. They illustrated how RT data can help understand the factors that contribute to variations in RT and possibly identify construct-irrelevant variances. While the impacts of test characteristics were more pronounced in the math test, no significant effects were observed at the medium and high test level for the reading test. Meanwhile, it was observed that test takers with English as a second language needed more time in both content areas and at different proficiency levels, compared with test takers with English as their first language.

The potential differentiating effects of content areas and cognitive demands may be observed in Parshall et al. (1994) and Halkitis and Jones (1996). The former study investigated the possible determinants of RT on a computer-adaptive college placement test of mathematics from the perspectives of both items and test takers. Their regression-based analysis indicated that none of the predicting variables, individually or combined, could explain the RT variability. Given that the computerized placement test was fairly new to the participants, Parshall et al. suspected that the novelty effect may have masked possible effects on RT. Parshall et al.'s findings are not consistent with those in Halkitis and Jones' research, which studied the relationship between test taker's RT and task characteristics in a national licensing examination in the U.S. By using regression analyses, Halkitis and Jones reported that three characteristics (item difficulty, discrimination, and item word counts) could explain 50.18% of the variance of the logarithm transformed RTs. More RT is needed in the case of high item discrimination, or long items, or difficult items. Similarly, Gvozdenko and Chambers (2007) employed a quasi-experimental design in their study on RT data in a university test of basic mathematics skills. Through content-expert judgment and self-reported strategy use from a cohort similar to the test taker population, Gvozdenko and Chambers were able to associate individual differences in RTs with their strategy use as well as differential cognitive demands in three parallel versions of the test in 2006. Their study indicates that RT data can be valuable in providing information about the cognitive load of test questions.

Test takers' use of RT can also be affected by their learning experiences and certain personal traits. Lasry, Watkins, Mazur, and Ibrahim (2013) focused on students' RT spent on conceptual questions in a university-level introductory physics course. Alongside the RT data, the students also indicated their confidence level for each item in a pre- and post-instruction design. In the pre-test, as expected, Lasry et al. found that students spent more time on the items that they answered incorrectly than on the ones that they answered correctly. At the end of the instruction period, the students spent more time on both the correct and the incorrect answers, a phenomenon that Lasry et al. attribute to the instruction; it appears to have taught



them to think more carefully about the question and the concepts being assessed. In addition, students' self-reported confidence levels were negatively correlated with their RT data. In other words, when students were less confident in their responses, they needed more time to process the item and decide on their responses.

Another study exemplifying the effects of test taker characteristics on response behavior is Hess et al. (2013). Hess et al. studied the RT behavior of test takers with various characteristics on a medical certification exam, which consists of two item formats, namely traditional multiple choice items and complex graphic-intensive multiple response items. Hess et al. were particularly interested in age effects but also considered the possible effects of test taker gender and ability upon their RT behavior. They reported that test takers' gender, ability, and age partially explained their differences in RT in the two item formats. In particular, they found that the older test takers proceeded more slowly through the multiple-choice items, gathering pace at a slower rate than other test takers. Hess et al. attributed this to changes in processing speed as people age; older test takers simply need longer to process information than younger test takers. Interestingly, however, the older test takers initially responded more quickly to the complex graphic-intensive multiple response items (although the difference between ages leveled out with additional items) perhaps because these items drew on clinical processes that all experienced physicians, regardless of age, have automatized.

In a study on the reading sections of an international language assessment, Lee and Haberman (2016) utilized both RT and test-taking pacing to explore test-taking behaviors. They found that test takers from different native countries (China, France, Germany, and Korea) adopted different time-allotment strategies and exhibited different patterns of test progression. In general, the Chinese and Korean test takers proceeded more slowly through the test than the French and German test takers. The Chinese and Korean test takers' response times were also more variable. When Lee and Haberman inspected the data more closely, they found that the Chinese and Korean test takers tended to move to the items before reading the input material while the French and German test takers tended to first read the reading passages before attempting the questions. They speculate that these test-taking behaviors affected their RT patterns and may be related to test preparation experience, test-taking strategies, and more importantly their reading speed.

Finally, Bergstrom, Gershon, and Lunz (1994) used hierarchical level models in a computer adaptive certification examination to explore the impact of test taker characteristics and item features on test takers' RT. They found that there were more within-person variance compared with between-person variance. More specifically, the results indicated that test anxiety was the only test taker trait that contributed to longer RTs, while item features such as item length, position of answer keys, and use of figures exerted significant impact on RTs.

## Has RT Lived Up Its Promise as Validity Evidence?

This chapter has described studies investigating the utility of RT in measurement research. The findings of these studies have been useful both for testing practice and for the development of testing theory. Nevertheless, we should take heed of Molenaar's (2015) caution that "the fact that the RTs are so easily available does not imply that they are useful" (p. 177). Most of the studies we have summarised use RT to infer the reasons for test taker behavior but the connection between the behavior (response time) and the reason for that behavior is still an open question. For instance, RT is used to identify rapid guessing behavior and inferences are drawn, such as low test taker motivation/effort or test taker pre-knowledge of the item. However, the RT information does not actually explain the cognitive processes that are involved in question-answering nor why they are used. It simply offers a measurable indicator.

Most of the RT studies reviewed here are exploratory in nature and have not been framed as validation studies, which are intended to be purposeful endeavors. In other words, validating score meaning has not been a primary purpose of these studies. Indeed, to date, RT data have not been treated as directly related to validity. From this point of view, the current status of research on RT does not reflect the potentially valuable role of RT in validity and therefore RT data have not yet fulfilled their promise as validity evidence. Of course, some of the findings from the aforementioned studies could be re-purposed or positioned as validity evidence, especially those that have shed light on response processes. In this sense, we see the great potential of RT data in future validation studies, especially if these studies were supported with stronger explanatory theories of response processes.

While the cognitive models of RT offer more explanatory power regarding response processes, they are very limiting in the sense that they are only applicable to relatively simple cognitive tasks. Possibly due to their inheritance of the typical perceptual-motor tasks in psychology lab studies, the research scenarios involving RT in educational assessment generally center on rather discrete or disjoint item formats, most of them being multiple-choice questions. With educational assessment embracing more complex tasks and innovative formats enabled by computer technologies, the current paradigm for RT research may fall short in accommodating these popular tasks. For example, contextualized or scenario-based tasks in the testing projects of mathematics or science, as well as productive tasks such as speaking and essay writing in language assessment, would require extended RT. In those situations, non-RT or preparation time may outweigh the observable RT with traceable actions. Understanding the cognitive processes that might occur over a relatively long span of time would call for a substantiation of non-RT and RT with proper theories.

## What Would Be Needed to Strengthen the Use of Response Time as Validity Evidence?

If RT data alone cannot adequately serve the purposes of validating test score interpretation, what would it take to do so? In our opinion, at least two approaches are relevant here: (a) connecting construct operationalization with (explanatory) cognitive theories, and (b) incorporating RT data with other process data.

In line with Zumbo's (2007, p. 52) call for explanatory rather than descriptive cognitive models of testing, the key to useful response time models is that they need to be explanatory and not just another set of descriptive models in psychophysical or cognitive terms rather than mathematical psychometric terms. It is important to note that not all response time models are explanatory. Put differently, a change of terminology in psychometrics from mathematical terminology to psychophysical or cognitive terminology is insufficient to claim true advances in gathering more meaningful and weighty validity evidence.

Most, if not all, cognitive process models are, at best, only capable of treating cognitive processes related to an isolated ability or trait (van der Maas et al., 2011). Many of the reviewed studies share one limitation in the processing of RT data: inadequacy of explanatory theory. In recent years, some explanatory models such as cognitive diagnosis models and explanatory IRT models have shown the possibility of conceptualizing an individual's mastery level as a development of multiple skills or abilities and associating test responses with the theory-based mastery levels (Gierl & Leighton, 2007; Gorin, 2006; Jang, 2009). Still, we feel a desperate need for more cognitive theories about response processes that will account for what happens during a given RT in both simple and complex tasks.

It is acknowledged that RT data only provide information about how much time test takers have spent on individual items and the same set of data alone could not "tell the full story about how test takers complete a test" (Lee & Haberman, 2016, p. 242). If we treat RT data as a temporal container, a number of other process data could be collected throughout the test-taking process delimited by RT intervals. As Messick (1989) has pointed out, RT used in "chronometric analysis" is more appropriate for the tasks that require relatively short time investment. Test-taking process information from other methods such as protocol analysis or verbal reports, cognitive interviews, and the analysis of eye-movement or keystroke logging should be used to facilitate understanding of test takers' cognitive processes, especially in cases of complex cognitive tasks (Messick, 1989; Zumbo & Chan, 2014). For example, eye-tracking data in educational contexts are particularly valuable for monitoring and understanding complex interactions and the cognitive processes of reading and graph viewing (Suvorov, 2015). In addition, keystroke logging techniques could be of tremendous importance in investigating the response processes that require text input, such as short essay tasks and open-ended questions (Leijten & Van Waes, 2013).

The RT literature in assessment contexts has demonstrated the potential utility of RT in assessment practices. It is tantalizing in its promise but in order to maximize

the usefulness of RT data for validation purposes, we will need to be equipped with better explanatory theories from cognitive psychology as well as from specific subject areas so that we can go beyond the descriptive approaches to processing RT data, together with other sources of process data. It is clear that we are currently in a state of anticipation; the potential of RT as validity evidence is closer than ever to being realised.

## 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.
- Bergstrom, B., Gershon, R., & Lunz, M. E. (1994, April). *Computerized adaptive testing exploring test taker response time using hierarchical linear modeling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Chan, S.-C., Lu, T.-S., & Tsai, R.-C. (2014). Incorporating RT to analyze test data with mixture structural equation modeling. *Psychological Testing, 61*, 463–488.
- Dennis, I., & Evans, J. S. B. T. (1996). The speed-error trade-off problem in psychometric testing. *British Journal of Psychology, 87*, 105–129. doi:10.1111/j.2044-8295.1996.tb02579.x.
- Dodonova, Y. A., & Dodonova, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence, 41*, 1–10. doi:10.1016/j.intell.2012.10.003.
- Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*, 655–670. doi:10.3102/1076998611422912.
- Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*, 525–543. doi:10.1177/0146621606295197.
- Fox, J.-P., Entink, R. K., & van der Linden, W. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software, 20*, 1–14. doi:10.18637/jss.v020.i07.
- Gierl, M. J., & Leighton, J. P. (2007). Defining cognitive diagnostic assessment in education. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 3–18). Cambridge, UK: Cambridge University Press.
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*, 21–35. doi:10.1111/j.1745-3992.2006.00076.x.
- Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.
- Gvozdenko, E., & Chambers, D. (2007). Beyond test accuracy: Benefits of measuring response time in computerised testing. *Australasian Journal of Educational Technology, 23*, 542–558. doi:10.14742/ajet.v23i4.1251.
- Halkitis, P. N., & Jones, J. P. (1996, April). *Estimating testing time: The effects of item characteristics on response latency*. Paper presented at the annual meeting of the American Educational Research Association, New York.
- Hess, B. J., Johnston, M. M., & Lipner, R. S. (2013). The impact of item format and test taker characteristics on response times. *International Journal of Testing, 13*, 295–313. doi:10.1080/15305058.2012.760098.
- Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*, 16–25. doi:10.1111/j.1745-3992.2001.tb00066.x.

- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*, 31–73. doi:[10.1177/0265532208097336](https://doi.org/10.1177/0265532208097336).
- Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods, 14*, 54–75. doi:[10.1037/a0014877](https://doi.org/10.1037/a0014877).
- Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*, 606–619. doi:[10.1177/0013164406294779](https://doi.org/10.1177/0013164406294779).
- Kahraman, N., Cuddy, M. M., & Clauser, B. E. (2013). Modeling pacing behavior and test speededness using latent growth curve models. *Applied Psychological Measurement, 37*, 343–360. doi:[10.1177/0146621613477236](https://doi.org/10.1177/0146621613477236).
- Lasry, N., Watkins, J., Mazur, E., & Ibrahim, A. (2013). Response times to conceptual questions. *American Journal of Physics, 81*, 703. doi:[10.1119/1.4812583](https://doi.org/10.1119/1.4812583).
- Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*, 359–379.
- Lee, Y.-H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing, 16*, 240–267. doi:[10.1080/15305058.2015.1085385](https://doi.org/10.1080/15305058.2015.1085385).
- Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*, 358–392. doi:[10.1177/0741088313491692](https://doi.org/10.1177/0741088313491692).
- Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*, 29–37. doi:[10.1111/j.1745-3992.2007.00106.x](https://doi.org/10.1111/j.1745-3992.2007.00106.x).
- 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* (pp. 313–319). New York, NY: Springer.
- Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*, 426–451. doi:[10.3102/1076998614559412](https://doi.org/10.3102/1076998614559412).
- Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement, 52*, 1–27. doi:[10.1111/jedm.12060](https://doi.org/10.1111/jedm.12060).
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.
- Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34*, 521–538. doi:[10.1177/0146621609355451](https://doi.org/10.1177/0146621609355451).
- Mislevy, R. J. (1989). Foundations of a new test theory. *ETS Research Report Series, 1982*(2), i-32.
- Molenaar, D. (2015). The value of response times in item response modeling. *Measurement: Interdisciplinary Research and Perspectives, 13*, 177–181. doi:[10.1080/15366367.2015.1105073](https://doi.org/10.1080/15366367.2015.1105073).
- Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research, 50*, 56–74. doi:[10.1080/00273171.2014.962684](https://doi.org/10.1080/00273171.2014.962684).
- Parshall, C. G., Mittelholtz, D. J., & Miller, T. R. (1994, April). *Response latency: An investigation into determinants of item-level timing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.
- Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice, 35*, 38–47. doi:[10.1111/emip.12102](https://doi.org/10.1111/emip.12102).
- Ranger, J., & Kuhn, J.-T. (2013). Analyzing response times in tests with rank correlation approaches. *Journal of Educational and Behavioral Statistics, 38*, 61–80. doi:[10.3102/1076998611431086](https://doi.org/10.3102/1076998611431086).
- Ranger, J., & Kuhn, J.-T. (2014). Testing fit of latent trait models for responses and response times in tests. *Psychological Test and Assessment Modeling, 56*, 382–404.

- Ranger, J., & Kuhn, J.-T. (2015). Modeling information accumulation in psychological tests using item response times. *Journal of Educational and Behavioral Statistics*, *40*, 274–306. doi:[10.3102/1076998615583903](https://doi.org/10.3102/1076998615583903).
- Ranger, J., & Kuhn, J.-T. (2016). A mixture proportional hazards model with random effects for response times in tests. *Educational and Psychological Measurement*, *76*, 562–586. doi:[10.1177/0013164415598347](https://doi.org/10.1177/0013164415598347).
- Ranger, J., Kuhn, J.-T., & Gaviria, J.-L. (2015). A race model for responses and response times in tests. *Psychometrika*, *80*, 791–810. doi:[10.1007/s11336-014-9427-8](https://doi.org/10.1007/s11336-014-9427-8).
- Ranger, J., & Ortner, T. M. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*, 128–148.
- Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 870–888. doi:[10.1037/a0034954](https://doi.org/10.1037/a0034954).
- Ratcliff, R., van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261–300.
- Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37–50. doi:[10.1016/j.intell.2014.10.003](https://doi.org/10.1016/j.intell.2014.10.003).
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232. doi:[10.1111/j.1745-3984.1997.tb00516.x](https://doi.org/10.1111/j.1745-3984.1997.tb00516.x).
- Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc..
- Siem, F. M. (1996). The use of response latencies to enhance self-report personality measures. *Military Psychology*, *8*, 15–27. doi:[10.1207/s15327876mp0801\\_2](https://doi.org/10.1207/s15327876mp0801_2).
- Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, *32*, 463–483. doi:[10.1177/0265532214562099](https://doi.org/10.1177/0265532214562099).
- Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179–203). New York, NY: Academic Press.
- Thomas, M. H. (2006). *Modeling differential pacing trajectories in high stakes computer adaptive testing using hierarchical linear modeling and structural equation modeling*. Unpublished doctoral dissertation. The University of North Carolina at Greensboro, Greensboro, NC.
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204. doi:[10.3102/10769986031002181](https://doi.org/10.3102/10769986031002181).
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. doi:[10.1007/s11336-006-1478-z](https://doi.org/10.1007/s11336-006-1478-z).
- van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272. doi:[10.1111/j.1745-3984.2009.00080.x](https://doi.org/10.1111/j.1745-3984.2009.00080.x).
- van der Linden, W. J. (2011). Setting time limits on tests. *Applied Psychological Measurement*, *35*, 183–199. doi:[10.1177/0146621610391648](https://doi.org/10.1177/0146621610391648).
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384. doi:[10.1007/s11336-007-9046-8](https://doi.org/10.1007/s11336-007-9046-8).
- van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.
- van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347. doi:[10.1177/0146621609349800](https://doi.org/10.1177/0146621609349800).
- van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195–210. doi:[10.1177/01466219922031329](https://doi.org/10.1177/01466219922031329).



- van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356. doi:[10.1037/a0022749](https://doi.org/10.1037/a0022749).
- Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, *60*, 385–402. doi:[10.1027/1618-3169/a000218](https://doi.org/10.1027/1618-3169/a000218).
- Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *The British Journal of Mathematical and Statistical Psychology*, *66*, 144–168. doi:[10.1111/j.2044-8317.2012.02045.x](https://doi.org/10.1111/j.2044-8317.2012.02045.x).
- Wang, T. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339. doi:[10.1177/0146621605275984](https://doi.org/10.1177/0146621605275984).
- Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85. doi:[10.1016/0001-6918\(77\)90012-9](https://doi.org/10.1016/0001-6918(77)90012-9).
- Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated test takers. *Journal of Computerized Adaptive Testing*, *2*, 1–17. doi:[10.7333/jcat.v2i0.30](https://doi.org/10.7333/jcat.v2i0.30).
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38. doi:[10.1111/j.1745-3984.2006.00002.x](https://doi.org/10.1111/j.1745-3984.2006.00002.x).
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of test taker motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183. doi:[10.1207/s15324818ame1802\\_2](https://doi.org/10.1207/s15324818ame1802_2).
- Yang, X. (2007). Methods of identifying individual guessers from item response data. *Educational and Psychological Measurement*, *67*, 745–764. doi:[10.1177/0013164406296978](https://doi.org/10.1177/0013164406296978).
- Zenisky, A. L., & Baldwin, P. (2006). *Using item response time data in test development and validation: Research with beginning computer users*, Center for educational assessment report No. 593. Amherst, MA: University of Massachusetts, School of Education.
- Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Psychometrics* (Vol. 26, pp. 45–79). The Netherlands/Amsterdam: Elsevier Science B.V..
- Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences*. New York, NY: Springer.

# Chapter 10

## Observing Testing Situations: Validation as Jazz

Bryan Maddox and Bruno D. Zumbo

### Introduction

In this chapter, we describe how observations of real-life testing situations can provide insights into test validation by focusing on response processes and test performance that are not easily captured by large-scale quantitative data or by conventional “think aloud” protocols. Think aloud protocols are considered by some to be the received method for investigating response processes from an individual cognitive perspective. In contrast, we consider real-life testing situations as distinctive social occasions that merit observation (Maddox, 2015). While testing situations reveal observable structures and patterns of behaviour, every performance is somewhat different. Like jazz, investigating the testing situation involves elements of improvisation. We see our task as to listen to those patterns and improvisations. That is, to hear music rather than noise.

The approach taken in this chapter is informed by Goffman’s assertion that “micro-analysis” of small-scale, face-to-face interactions is a useful mode and domain of social enquiry (e.g., Goffman, 1964, 1983). We illustrate this argument with short video-ethnographic transcripts from observations in Slovenia of the Survey of Adult Skills – the OECD Programme for the International Assessment of Adult Competencies (PIAAC).

---

B. Maddox (✉)

School of International Development, University of East Anglia, Norwich NR4 7TJ, UK

Laboratory of International Assessment Studies, Norwich, UK

e-mail: [B.Maddox@uea.ac.uk](mailto:B.Maddox@uea.ac.uk)

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: [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca)



We consider the significance of intimate, face-to-face interaction, as a starting point to investigate how variation in assessment performance is informed by, and a response to, the ecology of the testing situation.

By focusing on observations of interaction in face-to-face testing situations and the character of improvisations, we expand the set of information available to understand and explain response processes (see Zumbo, 2007a; Zumbo et al. 2015). However our aim is not simply to amplify individual differences in test behaviour. Instead, by observing the testing situation we hope to identify clues about the way the test is constructed, understood, and performed as a social occasion. This may include, for example, observation of interaction within wider social structures or social relations that inform and mediate assessment performance. These act as enabling conditions for the abductive explanation for variation in test performance. As such, our aim is guided by a contextualized Pragmatic form of abductive explanation as validation (Zumbo, 2007b, 2009). In terms of the process of validation (as opposed to validity, itself), the methods described herein work to establish and support the inference to the best explanation—i.e., validity itself; so that validity is the contextualized explanation, whereas the process of validation involves the myriad methods of psychometrics, including what we call “psychometric-ethnography” (Maddox, Zumbo, Tay-Lim, & Qu, 2015). Zumbo’s abductive approach to validation seeks the enabling conditions through which a claim about a person’s ability from test performance makes sense (Stone & Zumbo, 2016; Zumbo, 2007b, 2009).

We employ the rhetorical device of testing ‘in-vivo’ to capture the process of interaction and social embeddedness of the testing situation that inform and mediate individual response processes (Zumbo, 2015b). Although it may not be considered as construct relevant, such ecological information provides a potential explanation for variation in response processes rather than being considered as a source of pollution or cultural noise to be controlled and excluded. The contrasting idea is that assessment practice and explanation could somehow occur “in vitro,” as if isolated from its cultural and ecological setting, and sources of influence that occur in real-life operational contexts.

Testing agencies do, of course, have an obligation to identify and remove sources of bias and inequality in assessments that may arise from tacit cultural knowledge, inappropriate test constructs, materials, or administration (Hambleton, 2005). However, since it is neither desirable nor possible to remove all sources of construct irrelevant ecological sources of variation, we have an obligation to recognise and understand how those factors influence response processes (McNamara & Roever, 2006; Zumbo, 2015a). In so doing, we turn “noise” into “music.”

In discussing various metaphors for test validation such as stamp collection, chains of inference, or judicial and courtroom metaphors, Zumbo (2007b, p. 72) introduced “validation as jazz.” The principal tenets of validation as jazz is not only sound coming together, converting what some hear as noise to music, but also that the coming together is not necessarily scripted. All sorts of notes, chords, melodies, and styles come together (including, of course, improvisation that is particular to that one song or performance) in a creative way to move from noise disturbance to making music. To take the validation as jazz metaphor one step further, it involves

recognizing the context of testing as enabling conditions to understand and explain test performance in investigating response processes in validation as jazz sounds, songs, and styles that all come together to make music--there is no one methodology or script for test validation that can be applied in all measurement contexts.

The perspective on response processes that we adopt in this chapter therefore focuses on the dynamics of interaction in testing situations (e.g., see Maddox, 2015; McNamara, 1997), while recognising the potential for those interactions and responses to be influenced by larger-scale “off-stage” (Goffman, 1959) dimensions of the testing situation such as social institutions, social relations, norms, and beliefs that we might associate with “ecological” models of testing (e.g., Bronfenbrenner, 1994; Fox, 2003; Goffman, 1959; Maddox et al., 2015; Zumbo, 2007a, 2007b; Zumbo et al., 2015). Instead of viewing the ecological and interactive context as somehow inimical to the effective study of response processes, we consider it as the ideal setting to investigate test-taking behaviour.

The chapter is structured in the following way. *Theory and Method* describes the key approaches to the study of face-to-face testing situations. Informed by Goffman’s micro-analysis of social situations, we describe an approach to the study of testing situations that draws on research traditions in linguistic anthropology and conversation analysis. *Assessment Observed* illustrates the method with short transcripts from sequences of interaction in the Slovenian PIAAC assessment. The transcripts illustrate temporal, interactive, and affective characteristics of the testing situation. They illustrate the potential for studies of the testing situation to provide new insights into response processes. In the final part of the chapter, *Integrating Data*, we consider how micro-analytic insights from testing situations can inform, and be integrated into, validation practice.

## Theory and Method

The micro-analytic study of response processes in real-life (naturalistic) testing situations is a relatively new focus in the study of response processes. However, observations of face-to-face interaction have strong theoretical and methodological foundations in the interactionist tradition established by Goffman (1959, 1964, 1981, 1983) and subsequent development in Linguistic Anthropology (e.g., Duranti & Goodwin, 1992; Goodwin, 2007a, 2007b), Conversation Analysis (e.g., Shegloff, 1988; Shegloff & Sacks, 1973), and modern Gesture Studies (e.g., Goodwin, Cakaite & Goodwin, 2012; McNeill, 1992, 1985). As we suggest in this paper, these fields provide a body of literature that can readily be applied to the study of testing situations, and that supports wider inter-disciplinary analysis about the ecology of testing situations.

For Goffman (1959, 1964), social situations in different contexts involve the participants in culturally defined norms and obligations, with distinctive types of behaviour, talk, and interaction. We can apply Goffman’s insights to investigate testing situations as distinctive forms of social occasion (Maddox, 2015). Within the

testing situation, we can, for example, consider the rituals of test taking and the dynamics of test interaction and responses, observing test behaviour as it occurs in social and material settings. This not only offers a distinctive domain of enquiry, but also provides an empirical basis to investigate such phenomenon.

Testing situations vary considerably in their characteristics, and it is that variation that makes them a viable domain of investigation. However, they also have distinctive and regular features that support their study and comparison within settings and across cultures. There is not space here to provide an exhaustive description of the features of different kinds of testing situations or the variation that is observed across cultural contexts (e.g., see Maddox, 2014). Nevertheless, for the purposes of this chapter we can describe some of the features that inform such enquiry.

### *Institutional and Cultural Settings*

The social significance and meaning of a test is not principally derived from within the testing situation itself, but by broader institutional, political, and ideological arrangements relating to the purpose of the test and its consequences. Observations of testing situations provide the opportunity to consider how test responses are shaped by larger-scale social structures and social relations (Giddens, 1988), such as the household, school, community and nation. This might help to better understand and explain variation in test motivation and performance across different social contexts (e.g., Eklöf, 2010). This suggests a shift away from normative classifications of “low stakes” or “high stakes” assessment to allow such judgements to be informed by research.

Test respondents may have social ties and obligations to social institutions in ways that imply or generate feelings of obligations or consequence. This may be obligations toward family members, a sense of civic pride or nationalist competition with respondents in other regions or countries, a disagreement about the ideological purpose or intended use of the test, or a sense of animosity or a sense of mistrust with the institutions or individuals who administer the test. These ecological influences relate to different micro, meso, and macro scales of social structure (see Bronfenbrenner, 2005; Giddens, 1988), and involve contextual sources of variation that lie outside in-vitro notions of isolated individual cognition (e.g., see Bronfenbrenner, 1979, 2005; Fox, 2003; Zumbo et al., 2015).

### *Participation*

If we focus on the immediate context of the testing situation, including the social material setting of the test, we can examine the significance of social interaction—talk, gesture, emotion—in the assessment process. To make sense of such an interaction, we use Goodwin’s (2000, 2002, 2007a, 2007b) participation framework which

describes how individuals in face-to-face activities orient their bodies, talk and gesture toward each other, and demonstrate a “mutual orientation” to their shared task or activity:

A primordial site for the analysis of human language, cognition, and action consists of a situation in which multiple participants are attempting to carry out courses of action in concert with each other through talk while attending to both the larger activities in which their current actions are embedded and relevant phenomena in the world around them. (Goodwin, 2002, p. 2)

Testing situations frequently demonstrate such an interaction as the participants—respondents and interviewers—share a mutual orientation towards the test materials or the computer and the assessment tasks. This makes them ideal sites to examine the content and significance of verbal participation and gesture. Many testing situations are not undertaken in silence. This is certainly the case in the Slovenian PIAAC assessment described below, as we can observe considerable verbal interaction that takes place between the interviewer and respondent, and occasionally with other family members who may participate in the testing situation as ‘bystanders’ (on bystanders, see Goffman, 1964, 1981).

The presence of talk and gesture is particularly important in the observations of testing situations; as Duranti and Goodwin (1992) note, such interaction plays a role in the production of social context (i.e., the characteristics of interaction, a shared sense of what is going on and associated obligations). Talk and gesture in tests is therefore intimately connected with response processes. In contemporary gesture studies, McNeil (1985, 1992) argues that talk and gesture have a shared origin in social interaction. Goodwin et al. (2012) argue that gesture should be viewed as a public form of communication (i.e., often integral to talk), rather than an accidental expression of an inner state. Furthermore, as Goodwin (2007a, 2007b) notes, talk and gesture are often ‘environmentally coupled,’ in the sense that their meaning is tied to the material environment. In other words, the meaning of talk, or indeed of gesture, is not necessarily explicit without locating it in a material environment.

The implication of such studies is that, while transcripts of talk (e.g., in think aloud protocols) may provide meaningful information and feedback about how respondents understand and receive test items, “in-vivo” studies of interaction and behaviour in real-life testing situations locate talk and gesture within the testing environment, and within time and place.

### *Sequences of Interaction*

One of the characteristics that makes small-scale, micro-analytic studies of the testing situation a particularly useful domain of enquiry is the orderly, sequential characteristics of most test practices. The design of tests, and of test items, usually takes the form of predictable patterns of interaction. Proponents of Conversation Analysis are keen to point out the sequential characteristics of talk and gesture in social interaction (e.g., see Duranti & Goodwin, 1992; Goodwin, 2002; Shegloff, 1988;

Shegloff & Sacks, 1973). Those sequences of interaction—whether they are extended sequences of talk or short “adjacency pairs” (Shegloff & Sacks, 1973)—involve expression of a shared understanding about what is going on, what Goffman (1983) called felicity conditions. Those ideas also make an important contribution towards understanding the orientation of the respondent and interviewer to the test, that is, themes which have variously been described as respondent “engagement,” “will,” “motivation,” and “resilience.” These can be considered theoretically as a concern with “stance” (see Du Bois & Kärkkäinen, 2012; Goodwin et al., 2012). The stance or affective orientation of participants towards each other, and towards a focal activity is observed in the characteristics of interaction. In that sense, talk and gesture provide on-going, real-time, and “public” displays of stance (Goodwin et al., 2012) in the testing situation.

The structure of testing situations also involves distinctive sequential and temporal characteristics. There is usually an initial discussion about the test, its purpose and procedures. In the opening minutes, the “rules” of the test might be presented (e.g., certain expectations about the conduct of the test taker). The roles of respondent and interviewer may be defined, and any time requirements may be presented. In the case of written tests or computer-based assessment, as the test begins, the test taker may ask additional questions about the procedures for answering the particular test items as they begin to engage with the test. In the PIAAC assessment, respondents frequently asked questions during the assessment about procedures for answering particular items, and about the time duration of the test. This was complicated by the fact that, in the computer-based assessment, the interviewer did not know exactly how long the test would take.

Respondents in the PIAAC assessment sometimes commented in an evaluative sense about the test items or their performance. This included comments about their own performance on particular items, or their psychological state (e.g., fatigue, lack of concentration, boredom, or enjoyment). In most cases, verbal comments were accompanied by non-verbal gestures or facial expressions. In some cases, this provided a stream of communication (particularly facial expressions and non-verbal gestures) as they completed the assessment.

In the final stages of the assessment, as the test is completed, there are certain patterns and routines. In the case of the PIAAC assessment in Slovenia, this usually afforded opportunities for some informal discussion about the test, and the respondents’ performance and feelings. There were often evaluative comments about the experience of taking the test, and some relief about its completion. This included opportunities for some discussion about how the test data would be used (i.e., the test purpose).

The sequential structure of the test, and the precise recording of item responses (and in the case of the PIAAC assessment, log file data on response times and key-strokes), mean that interactive data and observations from the testing situation (talk, gesture, and facial expressions) can be integrated into the larger data on test response processes. This enables analysis of sequences of behaviour and detailed analysis of respondent engagement on particular items. Furthermore, observations of interaction in the testing situation can integrate analysis about the role of the interviewer in that process.

Interaction, talk, and gesture can therefore be studied as a coherent set of data that is intimately connected with the process of test taking and the ecological setting of assessment. The in-vivo perspective locates test items and response processes—with their various demands, and their sequences of action—within a more holistic ecology of the test as a social occasion. That involves dimensions of social interaction and improvisation, with endogenous spatial, temporal, affective, cultural, material, and institutional characteristics that exceed the conventional parameters of test design. This introduces multiple sources of information and meaning that influence response processes—what we might consider using the metaphor of jazz, as multiple melodies that combine in the testing situation. In the section below, we illustrate these arguments in relation to empirical data from the Slovenian PIAAC assessment and consider their implications.

## Illustrative Examples

The examples below use “video-ethnographic” methods to investigate interaction and response in the OECD PIAAC Assessment conducted in Slovenia in 2014 (PIAAC Round 2). PIAAC is a test of the skills of adults in literacy, numeracy, and problem solving in technology rich environments. The assessments are delivered in people’s homes either through paper based testing or on a laptop with a multi-stage computer-adaptive test (CAT) design (OECD, 2013; Yamamoto, 2011).

### *Example 1. Interaction in Assessment*

In this sequence, the respondent has completed the background questionnaire and has just started the computer-based assessment. In this testing situation,<sup>1</sup> the interviewer (I) sat adjacent to the respondent (R; note: this was not always the case in the assessments observed). The researcher sat opposite the interviewer. The small video camera rested on the Table.

- I: We’ll arrange it this way ... We’ll place it here, [so you’re more comfortable. ((while she is talking, the interviewer is arranging the laptop and mouse))
- R: [I see, it’s okay. That it.  
So we are here and go down here ((looking at the screen)), I see. Good. Let’s move on.  
(R looks up from the screen directly at the interviewer)).

---

<sup>1</sup>In the transcription, I = interviewer (test administrator), R = respondent (test taker), [square bracket indicates overlapping speech, and ((double parentheses indicate descriptions of what is going on.

- R: Will you do the talking?  
 ((R gestures as if typing in the air, then raised hand to indicate no)).
- I: No, no, I'm not allowed to do anything.
- R: I just move on? [Yes, yes, yes, yes yes.
- I: [You work independently, yes. That's it. ((R is looking at the screen, and the Interviewer has leaned over momentarily to see the screen, i.e., shared gaze at the screen))
- R: What about, I mean, what now? I've done this already. Choose a month—and I chose October—now what? Why in fact it again?
- I: Um, um, choose May. ((the interviewer has leaned over to see the screen)).
- R: I see! I did [no ((when the respondent realises what she has done she opens her eyes wide and gestures with her hand at 45 degrees to the screen. The interviewer leans over again to see the screen)).



- I: [The instructions are always at the top.
- R: I haven't read this at all, I read only 'select a month' and I chose the month we're in! I blew it! ((while saying this the respondent gestures indicating that she did something wrong, pointing at the screen, momentarily covering her face, and then looks directly at the interviewer and smiles)).
- I: The instructions are always at the top. ((the interviewer leans over to see the screen and *points* to the top of the screen)).



- R: Yes, yes, I have to take a look.  
 I: Nothing works with the ‘enter’ ((the interviewer points to the enter key)), it always goes [here ((she points to the section on the screen and smiles)).  
 R: [Yes yes yes, I [understand.  
 I: [OK. ((the respondent points her hand toward the screen, and nods with a serious face. The interviewer also nods then moves away, and the respondent continues with the assessment)).

The transcript displays the characteristics of interaction in testing situations. We see a sequence of talk and gesture that reflects the shared orientation to the assessment task. We see from the sequence above that the ‘context’ of the testing situation is produced, in part, through interaction between the interviewer and respondent (i.e., it establishes certain roles and boundaries). The focus of that interaction is almost always the computer screen, with only occasional interruptions by “off-screen” discussions that are marked by changes in body posture and gaze (e.g., direct eye contact). This illustrates the dynamic nature of the testing situation involving interaction among the respondent, computer, and interviewer.

The in-vivo perspective that we advocate recognises the significance of that interaction in the performance of assessment. That is, it does not consider the presence and role of the interviewer as peripheral to the assessment process, or as a source of measurement error (i.e., an “interviewer effect”). Instead, we can observe and consider the dynamics of joint participation of interviewer and respondent in the assessment task.

In the example, the active part played by the interviewer is somewhat at odds with her statement, “*I’m not allowed to do anything.*” The interviewer’s management of the testing situation reflects a sophisticated awareness of not only the tasks required by the assessment but also the affective, emotional orientation of the respondent. Much of the agency in the assessment is attributed to the computer. However, the interviewer, as a ratified bystander (Goffman, 1964), has retained an important role as a third-party, relating to communication of test taking procedures, procedural problem-solving, and emotional support.

### ***Example 2. Assessment in Time and Place***

The second example is from an assessment that took place in the evening in the respondents’ home. The respondent explained at the start that he was tired after a long day at work. During the assessment, the respondent’s young child waited in the next room with occasional forays past her father to obtain snacks from the kitchen. The respondent was well engaged with the assessment tasks, but also showed signs of tiredness in his facial expressions, suggesting that he was finding it hard to maintain his concentration. The sequence below comes at 35 min into the assessment.

((The respondent makes a long and marked sigh with a long puff out of breath and an exaggerated puffing out of his cheeks. The blowing out breath lasts for 9 s)).





((The respondent turns in a slow and deliberate manner while also lifting his coffee cup to look directly at the interviewer as she looks directly at him with a slightly worried expression on her face. His facial expression and tone is slightly accusatory, though not unpleasant)).

- R: Can you tell me how far we are? ((He smiles slightly as he asks)).  
 I: No. ((The interviewer says with a slight shake of her head)).  
 R: Ah. ((He smiles and turns away slightly as if to take a drink from his coffee, with a large smile on his face)).  
 I: You still have some exercises. You are over half, at two thirds. I'm speaking from [experience. ((As she speaks the respondent takes a sip of coffee and puts his cup down then smiles and nods)).  
 R: [yeah. ((As she speaks, the interviewer gestures with her hands left and right, makes an apologetic smile and opens her eyes wide and her head slightly bowed and tilted to one side. The respondent smiles in response, takes a sip of coffee and puts his cup down)).  
 I: [But I cannot influence the computer's selection of exercises for you. So that ... there can be a slight deviation. ((As the interviewer explains the respondent turns back to face the computer and continues with the assessment. He reads the item text indicating that he is reading by moving his lips)).

The sequence above was initiated nonverbally through the long sigh and facial expression. That illustrates the phenomenon of response cry that is discussed by Goffman (1981). Response cries in assessments were usually a precursor to verbal interaction. Like in this example, interactions in the PIAAC assessments observed frequently related to the length of time that the assessment was taking and the associated expression of boredom or fatigue.

The example shows how the context of the testing situation (to paraphrase Malinowski, 1923) impacts on assessment performance (i.e., tiredness after work, and responsibilities for family members). This was often the case for parents with young children – with childcare introducing a distinctive ecological dimension into

the assessment. In this case, the “in-vivo” perspective identifies characteristics of the testing situation that may have shaped response processes. These extend beyond face-to-face interaction to include micro and meso dimensions of the testing situation—that is, of the household, family, and employment (Bronfenbrenner, 1994; Zumbo et al., 2015).

Like the first example, this transcript highlights the importance of interaction. That is, the interviewer manages the situation at an emotional (affective) level (e.g., she is concerned about how respondent fatigue might impact on his engagement and performance). Here then, we become aware of the links between the micro-ecological ‘huddle’ (Goffman, 1964, p. 135) involving the respondent, interviewer, and the computer, and the wider meso and macro ecologies in which it is located—that is, the world of work, fatigue, and child-care responsibilities, and obligations of citizenship to participate in ‘low stakes’ assessment.

## Integrating Data

By introducing the metaphor of validation as jazz, and the idea of in-vivo observation, we consciously depart from routine procedures of validation practice, and question the idea of assessment context as a source of noise and pollution. Instead, we view the testing situation as a legitimate source of ecological and interactive information about response processes.

What implications do these transcripts have for the analysis of variation in response processes? In the first example, we can see that familiarity with the computer is one potential factor that will shape performance. This is evident from the interaction, and may also be observed in the computer-generated log files (i.e., on response processes and response timing). We can also see how the skills of the interviewer help the respondent to feel comfortable with the use of the computer and the format of the assessment tasks. The second example demonstrates the importance of interaction in assessment—and introduces wider ecological factors that may influence response processes. These observations provide in-vivo perspectives about how factors that are not necessarily relevant to the construct or trait under investigation can influence assessment performance.

Our view is that data obtained in real-life testing situations should not be viewed as chaotic and unruly. Rather, our task is to carefully listen to the unexpected melodies that such observations produce (e.g., the harmony of the interviewer working in collaboration with the computer, or the dissonance of test engagement with child care responsibilities and its impacts on time). These new sources of information, we suggest, can reasonably be integrated into validation practice. In considering how to do so, we would like to make the following three points.

Firstly, in drawing attention to the dynamics of the testing situation, our intention is *not* to undermine or debunk conventional approaches to assessment. We do not consider the “discovery” of interaction, social context, or the socio-political dimensions of testing as necessarily involving a threat to the validity of large-scale assessments.

However neglecting those ecological dimensions of the testing situation is increasingly viewed as problematic (e.g., Fox, 2003; McNamara, 2007; McNamara & Roever, 2006). Our aim (informed by Zumbo's "Third Generation" DIF analysis) is therefore to identify and explain sources of variation in test response processes that are endogenous to the testing situation and that lie outside individual notions of cognition (Zumbo et al., 2015).

Second, we *not* arguing that the microanalysis of interaction in face-to-face testing situations should, in any way, displace or usurp the role of large-scale data and psychometric analysis. Instead, we view such information as an opportunity to enrich the information base available in the analysis of response processes and assessment validity. Our aim is not to pit qualitative against quantitative, or to somehow suggest that large-scale enumerative approaches are invalid. We do not seek to privilege one over the other, and have developed a combined process of ethnographic-psychometric enquiry (see Maddox et al., 2015).

Finally, we are *not* suggesting that observations of the testing situation support a case for "business as usual." That is, micro-analytical studies of interaction and the testing situation do not simply extend the information base available for the analysis of assessment response processes. They inform theoretical and methodological questions about how the 'ecological' dimensions of each testing situation (i.e., its micro, meso, and macro characteristics) influence response processes and, more broadly, the "performance" of the testing situation as it occurs in real-life contexts—the perspective that we describe as *in-vivo* (Zumbo, 2015a, 2015b). In essence, this gives us many of the enabling conditions for validity as an abductive contextualized view of explanation. The challenge of integrating data is therefore about more than simply putting observations of testing situations at the service of conventional forms of validity practice.

## Conclusions

We have argued that the improvisations and contextual sources of variation observed in testing situations can be viewed as music rather than as noise. The idea of validation as jazz (following Zumbo, 2007b) involves careful listening and appreciation of the feedback offered by the multiple and complex dimensions of real-life testing situations. This is illustrated, in this chapter, through the micro-analytic observation and analysis of assessment interaction (Maddox, 2015). The approach is supported conceptually by the social situation (Goffman, 1964), and sequences of talk (Shegloff & Sacks, 1973) as units of analysis, and the by notion of 'in-vivo' observation of testing situations (Zumbo, 2015a, b) that we elaborate on in the chapter.

As we have demonstrated, naturalistic observations of interaction in testing situations can identify multiple endogenous sources of variation in test response processes. The status and significance of such variation becomes a question for validity analysis to consider. This supports an abductive and Pragmatic approach (Stone & Zumbo, 2016) to investigating test response processes, working between micro and macro scales of data analysis to provide contextualised and interactive understand-

ing and explanations of test performance (see Maddox, 2015; Maddox et al., 2015). In so doing, we shift the methodological focus from routine procedures of validation practice to a propensity to improvise and interact with feedback from real-life testing situations.

## References

- Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.
- Bronfenbrenner, U. (1994). Ecological models of human development. In T. Huston & T. N. Postlethwaith (Eds.), *International encyclopedia of education* (Vol. 3, 2nd ed., pp. 1643–1647). New York: Elsevier Science.
- Bronfenbrenner, U. (2005). *Making human beings human: Bioecological perspectives on human development*. Thousand Oaks, CA: Sage.
- Du Bois, J., & Kärkkäinen, E. (2012). Staking a stance on emotion: Affect, sequence and intersubjectivity in dialogic interaction. *Text and Talk*, 32(4), 433–451.
- Duranti, A., & Goodwin, C. (Eds.). (1992). *Rethinking context: Language as an interactive phenomenon* (pp. 1–42). Cambridge, UK: Cambridge University Press.
- Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, 17(4), 345–356.
- Fox, J. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing*, 3(1), 21–48.
- Giddens, A. (1988). Goffman as a systematic social theorist. In P. Drew & A. Wootton (Eds.), *Erving Goffman: Exploring the interaction order* (pp. 25–279). Oxford, UK: Polity Press.
- Goffman, E. (1959). *The presentation of self in everyday life*. Garden City, NJ: Doubleday.
- Goffman, E. (1964). The neglected situation. *American Anthropologist*, 66(6), 133–136.
- Goffman, E. (1981). *Forms of talk*. Philadelphia: University of Pennsylvania Press.
- Goffman, E. (1983). Felicity's condition. *American Journal of Sociology*, 89(5), 1–53.
- Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10), 1489–1522.
- Goodwin, C. (2002). Time in action. *Current Anthropology*, 43(S4), 1–53.
- Goodwin, C. (2007a). Participation, stance and effect in the organisation of activities. *Discourse & Society*, 18(1), 53–73.
- Goodwin, C. (2007b). Environmentally coupled gestures. In S. Duncan, J. Cassell, & E. Levy (Eds.), *Gesture and the dynamic dimension of language* (pp. 195–212). Amsterdam: John Benjamin's.
- Goodwin, M., Cakaite, A., & Goodwin, C. (2012). Emotion as stance. In M. Leena Sorjonen & A. Perakyla (Eds.), *Emotion in interaction* (pp. 16–41). Oxford, UK: Oxford University Press.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, R. P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Lawrence Erlbaum Associates.
- Maddox, B. (2014). Globalising assessment: An ethnography of literacy assessment, camels and fast food in the Mongolian Gobi. *Comparative Education*, 50(4), 474–489.
- Maddox, B. (2015). The neglected situation: Assessment performance and interaction in context. *Assessment in Education*, 22(4), 427–443.
- Maddox, B., Zumbo, B. D., Tay-Lim, B. S.-H., & Demin Qu, I. (2015). An anthropologist among the psychometricians: Assessment events, ethnography and DIF in the Mongolian Gobi. *International Journal of Testing*, 15(4), 291–309.

- Malinowski, B. (1923). The problem of meaning in primitive languages. In C. K. Ogden & I. A. Richards (Eds.), *The meaning of meaning* (pp. 296–336). New York: Harcourt.
- McNamara, T. F. (1997). “Interaction” in second language performance assessment: Whose performance? *Applied Linguistics*, 18(4), 446–466.
- McNamara, T. F. (2007). Language testing: A question of context. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 131–137). Ottawa, ON: University of Ottawa Press.
- McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, UK: Blackwell.
- McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, 92(3), 350–371.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: The University of Chicago Press.
- Organization for Economic Cooperation and Development. (2013). *Technical report of the survey of adult skills (PIAAC)*. Paris: Author.
- Schegloff, E. A. (1988). Goffman and the analysis of conversation. In P. Drew & A. Wootton (Eds.), *Erving Goffman: Exploring the interaction order* (pp. 89–135). Oxford, UK: Polity Press.
- Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica*, 8(4), 289–327.
- Stone, J., & Zumbo, B. D. (2016). Validity as a pragmatist project: A global concern with local application. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice* (pp. 555–573). Newcastle, UK: Cambridge Scholars.
- Yamamoto, K. (2011, September). *Implementation of CBT in the PIAAC field test and CAT in the PIAAC*. In: 20th seminar report at the Centre for Research on Educational Testing, Tokyo.
- Zumbo, B. D. (2007a). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223–233.
- Zumbo, B. D. (2007b). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Psychometrics* (Vol. 26, pp. 45–79). Amsterdam: Elsevier Science 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, NC: IAP – Information Age.
- Zumbo, B. D. (2015a). *Tides, rips, and eerie calm at the confluence of data uses, consequences, and validity*. Plenary address, “The Production of Data in International Assessments,” research conference organized by the Laboratory of International Assessment Studies, Economic and Social Research Council (ESRC), University of East Anglia, Norwich, UK. URL: <https://youtu.be/ahfbnLgUR5E>
- Zumbo, B. D. (2015b). *Consequences, side effects and the ecology of testing: Keys to considering assessment ‘in vivo’*. Keynote address, the annual meeting of the Association for Educational Assessment – Europe (AEA-Europe), Glasgow, Scotland. URL: <https://youtu.be/0L6Lr2BzuSQ>
- Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Astivia, O. L. O., & Ark, T. K. (2015). A methodology for Zumbo’s third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, 12(1), 136–151.

# Chapter 11

## A Rationale for and Demonstration of the Use of DIF and Mixed Methods

José-Luis Padilla and Isabel Benítez

Understanding group differences in item responses can provide insight into item response processes as test validation (Padilla & Benitez, 2014; Zumbo, 2007, 2009). In this light, as Zumbo (2007) notes, research into differential item functioning (DIF) can serve to provide a lens to item responding and validation by helping to establish for whom the test or item score inferences are valid, and for whom they are not. Few research problems have been so present through the history of test theory, and received so much interest from professionals, researchers, and testing organizations, as have item bias and DIF. Since it was first systematically addressed in the 1960s, the historical changes in how item bias and DIF have been understood and conceptualized can be traced following the series of releases of the *Standards for Educational and Psychological Testing*, published by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council of Measurement in Education (NCME). From the 1974 *Standards* wherein the problem of item bias was first introduced stemming from its serious social consequences through to the latest 2014 *Standards* wherein it is mainly treated as a fairness issue, such evolution has been inextricably linked to the history of validity theory.

The evolution of the conceptualization, analysis, and statistical methods of DIF is commonly interpreted by the “generations of DIF” concept proposed by Zumbo (2007). Although the author warned explicitly against interpretations in terms of historical periods or lineal progress, wondering what aims DIF research has achieved is hard to avoid. In the transition to the third generation when the article was

---

J.-L. Padilla (✉)  
University of Granada, 18071 Granada, Spain  
e-mail: [jpadilla@ugr.es](mailto:jpadilla@ugr.es)

I. Benítez  
Universidad Loyola Andalucía, 41014 Sevilla, Spain  
e-mail: [ibenitez.baena@gmail.com](mailto:ibenitez.baena@gmail.com)

published, DIF theorizing was characterized as conceiving DIF as a result of the item characteristics and/or “testing situation factors” that are not relevant to the intended construct. Bringing testing situation in a systematic way to DIF and item bias research is in the core of the present chapter. Mixed methods (MM) DIF and items bias studies can help not just in understanding DIF results but also to know why item bias occurs looking at the contextual and personal variables (e.g., cognition, attitudes, social location, etc.), traditionally ignored in DIF research.

In the chapter, we introduce MM research and the main characteristic of MM studies. We pay special attention to what defines a true MM study: the so-called “integration challenge,” commonly symbolized by the intriguing equation for researchers with a quantitative background “ $1 + 1 = 3$ .” The equation intends to convey the idea that a MM study is more than just put together a qualitative and a quantitative part—its whole being greater than the simple sum. Secondly, we posit a general framework to encourage professionals and researchers to conduct DIF MM studies that cast an eye to response processes and hence test validation. The main phases and the most appropriate ways of integration for DIF research are described. Finally, a research case study is thoroughly described as an example of a MM DIF study to illustrate the general framework. Throughout the chapter, we also present studies to illustrate the main ideas and proposals.

Due to the proliferation of overlapping labels (e.g., mixed method, mixed methods, mixed methodology, quantitative and qualitative research, etc.) and that MM research is quite a new approach in psychometrics, it is difficult to find examples of psychometric studies planned and conducted purposefully as MM research. To accommodate for this, we find ourselves periodically illustrating MM characteristics with studies that could be a MM study if authors had planned and conducted them following principles outlined in the chapter.

Before introducing MM research for DIF, a question that surely arises for researchers and professionals is, why do we need a new methodological framework, especially considered how far DIF research has gone, and how many DIF methods have been developed over the last 50 years of DIF theorizing and “praxis?” To address this question, the next section presents arguments in favor of incorporating MM research within the methodological approaches available for conducting DIF and item bias validation studies.

## Why Conduct a Mixed Methods DIF Study?

We would dare say that to include an agreed upon definition of DIF in the introduction of articles and chapters on DIF methods has turned out to be a kind of a convention or rule of protocol. The prototypical definition proposed by Millsap and Everson (1993) is, “difference in the functioning of ... an item among groups that are matched on the attribute measured by the ... item” (p. 298).

Willing to take risks, we would say that the treatment of DIF in the context of research on equivalence when measurements of different linguistics and/or cultural



groups are compared, has also become routine. DIF has an undisputed role in one of the most cited classifications of the bias level: (a) construct bias, occurs when the construct measured is not identical across cultural groups defined, (b) method bias, appears if samples are incomparable or problems related to the instruments characteristics exists, and c) item bias, which refers to distortions at item level (e.g., van Vijver & Leung, 2011). So the undisputed role that the statistical analyses of DIF to detect item bias is considered mandatory along with the study of other sources of bias, before affirming one of the groups has more or less amount of the variable measured than others.

Both in a monolingual research context and cross-lingual and cultural context, DIF is understood with slight nuances as a threat to the validity of score interpretations. In a monolingual DIF research context, the idea that any test or item parameter that is different between test takers from different subpopulation groups with similar overall ability may be a threat to validity is commonly accepted. Of course, the traditional warning that a distinction between DIF and item bias should be kept in mind is automatically sent. In the cross-lingual and cultural DIF research context, there is also a similar wide consensus that the presence of DIF threatens the validity of cross-cultural comparisons potentially invalidating any comparative interpretation based on total-score differences, or any group statistics based on them due to the lack of equivalence.

Together with the above shared views about the definition and relevance of DIF in monolingual and cross-lingual and cultural research, DIF researchers have provided a great arsenal of very powerful statistical methods to detect DIF.

### ***A Great Arsenal of Statistical Methods but Few Consolidated Results Investigating DIF Causes***

There is no doubt about how intensive psychometricians and statisticians have worked for the past three decades to develop statistical tests, effect size measures and criteria not just to flag items with DIF, but also to help users in making decisions on whether to keep or remove DIF items from tests. For recent state-of-the-art reviews on DIF research, readers should refer to the reviews by Hidalgo and Gómez-Benito (2010), Sireci and Rios (2013), and Zumbo (2007).

Describing the status reached at the “second generation” of DIF study, Zumbo (2007) presented the three frameworks for classifying groups of DIF statistical models on which most of DIF researchers could agree: (a) modeling item responses via contingency tables and/or regression models, (b) item response theory (IRT), and (c) multidimensional models. As Zumbo noted, the third framework of DIF statistical models was considered by some as particularly promising in the search for an explanation of why DIF occurs. The multidimensional models assume that some characteristic of the test item is not relevant to the underlying ability of interest could cause DIF, and that all tests are to some extent multidimensional. These principles would have allowed researchers to integrate DIF statistics with the powerful validity concept of “construct representation” (Embretson, 1983), which

includes as threats to validity, “construct underrepresentation” and “construct-irrelevant variance” (Messick, 1989). From our view, research on the causes of DIF has not exploited sufficiently the theoretical opportunity of the multidimensional framework until recently, when new DIF models have been developed.

Decades of DIF research have not provided an adequate understanding of the sources of DIF (e.g., Ferne, & Rupp, 2007; Penfield, 2010; Zumbo et al. 2015). This may be due to (a) the fact that routine performance of DIF analysis is sometimes motivated by legal and political reasons, and (b) an implicit acknowledgement of DIF and item bias seen as measurement flaws and problems that will always be there, and hence a willingness to either ignore it or accommodate for it.

We posit four reasons to explain the slow progress in the investigation into the causes of DIF and item bias. The first reason is the commonly observed lack of replicability or the incomplete convergence between statistical DIF results (Hidalgo & Gómez-Benito, 2010; Millsap & Everson, 1993). The second reason is the inability of traditional methods to provide information concerning the nature and location of DIF effects (Penfield, 2010). The third reason is the absence of a clear inclusion of DIF in a theoretical validity framework. The clear understanding of DIF as a validity issue would have posited DIF causes as the research questions for validation studies. And finally, researchers interested in finding DIF causes need a comprehensive methodological framework that can rigorously combine results and findings from quite different research methods. The following section further develops these reasons and presents our proposals to improve our understanding of DIF and item bias.

### ***The Lack of a Methodological Framework to Combine Traditional DIF Methods***

Hambleton (2006), in a commentary paper for the articles in a special issue of the journal *Medical Care* on differential item functioning (DIF) and factorial invariance, provided a perspective on how DIF research should be performed in practice. Among his suggestions, Hambleton refuted the myth that reviewers involved in judgmental reviews cannot be trained to point out aspects of the items or the instrument that are likely to impact differently on actual item performance of subgroups of the population. Indeed, not only reviewers can be trained, but also judgmental review would be much more efficient if clear questions to identify potential bias are included, such as the following:

- (1) Will the content of the item or statement be different or unfamiliar to individuals in designated subgroups of interest?...
- (2) Does the item or statement contain words that may have different or unfamiliar meaning for individuals in designated subgroups of interest? (p. s138).

Resorting to Hambleton’s (2006) recommendations on how to improve judgmental review is appropriate for at least two reasons. First, the recommendations highlight the importance of posing relevant questions to the reviewers, and second,

judgmental review is perhaps the DIF method more frequently combined with DIF statistics to interpret DIF results. Unfortunately, quite often judgmental reviews with the format of expert judgement, expert appraisal, panel of experts, or expert evaluation, etc., have been mixed with DIF statistics without a systematic plan.

In the last decade, more complex designs have been developed for combining results from different DIF methods, including judgmental review, in a more productive way (Benítez, Padilla, Hidalgo, & Sireci, 2015; Elosua & López-Jauregui, 2007; Ercikan et al., 2010). However, DIF studies combining judgmental reviews and DIF statistics are still far from providing a coherent picture of reasons for DIF.

There is clearly a need for a methodological framework that allows researchers to combine data from different methods taking advantage of their strengths and overcoming their respective weaknesses. This need is even more evident given the new DIF methods on the horizon such as those described in Zumbo's (2007) third generation DIF models. These third generation methods include consideration of the testing situation factors for items flagged as DIF or item bias. Investigation in to the extent that situational variables such as classroom size, socioeconomic status, teaching practices, parental styles, etc., will provide different kind of qualitative and quantitative data that needs to be combined with typical DIF results. Perhaps one of the most convincing arguments in favor of a comprehensive methodological framework comes from the analysis of one of the newest DIF methods: the use of the latent class logistic regression proposed by Zumbo et al. (2015). This method is motivated by the ecological model of item responding (Zumbo & Gelin, 2005) and the pragmatic and contextualized view of validity (Zumbo, 2009). As Zumbo et al. (2015) note, several layers of ecological variables can simultaneously effect item responding. These layers include, but are not limited to, not just the typical item characteristics seen in DIF studies, but also person characteristics, teachers, classroom, school context, family, community, etc. As such, this an explicit recognition of "neither the test taker nor the cognitive processes in item responding are isolated in a vacuum. Instead, test takers bring their social and cultural present and history to test taking" (Zumbo et al., 2015, p. 140). To include personal, contextual, social, and cultural factors in DIF research requires quantitative and qualitative data provided by very different methods, and as a consequence, a methodological framework that helps researchers to conduct such DIF studies.

### *Validity Claims for Integrations of Different Kinds of Data*

At this point, it is worth noting that a comprehensive methodological framework within which to perform DIF validation studies necessitates a contemporary view of validity and validation methods. It is beyond of the scope of the chapter to summarize the current status of validity theory; however, authoritative references to the recent theoretical work on validity theory and validation practices include Kane (2013), Sireci (2012), and Zumbo (2009). In broad strokes, the wide consensus about validity includes that (1) it belongs to the "entitled" inferences and

interpretations for the use of the test, (2) it is not a characteristic of the test or questionnaire, (3) it is a unitary concept, and (4) it is an evaluative judgment. A central idea for the current chapter is differentiating between the concepts of “validity” and “validation” as articulated by Zumbo (2009) in his pragmatic and contextualized view of validity as explanation of test score variation. Zumbo states that “validity is the explanation for the test score variation, and validation is the process of developing and testing the explanation” (p. 66). Beside this general consensus, the needs and claims for integration of different kind of data can be found in the 2015 keynote address to the Association for Educational Assessment – Europe about validity and validation (Zumbo, 2015).

In addition, the latest *Standards* (AERA, APA, & NCME, 2014) state that “validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11). Therefore, the *Standards* rely on the five sources of validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. Since there is also a clear acknowledgment that validity is a unitary concept, the *Standards* make an explicit call for integration of evidence by stating, “A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existence evidence and theory support the intended interpretation of test scores for specific uses” (p. 21, AERA et al., 2014). Given that “at some point validation evidence allow for a *summary judgment*...” (p. 22, AERA et al., 2014), mixed methods (MM) designs can make it easier to integrate evidence coming from different sources of validity and validation methods. Similar analysis can be done with validity frameworks close to the *Standards*. Both the so called “deconstructed” approach to validation by Sireci (2012), and the argument-based approach to validation (e.g., Kane, 2013), call for the integration of different kinds of data at some point of the validation processes. Finally, it would be difficult to find more explicit claims for integration than those made by Zumbo (2009) on presenting the core elements of the “integrative cognitive judgment of validity and the process of validation: validity, psychometrics, social consequence and matters of utility” (p. 69).

It is enough to understand DIF as a validity issue and perform research on DIF causes as *integrative* validation studies that will require, as Zumbo (2009) states, “consideration of the statistical methods, as well as the psychological and more qualitative methods of psychometrics, to establish and support the inference to the explanation” (p. 70) to recognize that MM DIF studies are knocking on the door.

## MM Research on DIF Causes

The contents of this section respond to two general questions: What is MM research? And what can MM studies provide research on DIF causes? We hope that assessment specialists and DIF researchers who are interested in deepening their understanding of DIF in mono- and multi-lingual or cultural testing recognize that the key

point in a true MM DIF validation study is integration. The more attention to integration they pay through the different phases of the study, the more solid integrative qualitative and quantitative evidence they will obtain to explain DIF and improve the validity of their test score interpretations.

MM research for some advocates represents an additional research paradigm, the third paradigm to distinguish it from the two established paradigms: quantitative and qualitative research (Gorard & Simon, 2010). A systematic introduction to MM research is beyond the scope of the chapter and likely unnecessary given the easy access to “classic” books published in the last few decades by, for example, Creswell and Plano Clark (2011), as well as Tashakkori and Teddlie (2003).

### ***Origins, Theoretical Foundations and Definition of MM Research***

When readers hear of the term “mixed research (MR),” probably the first idea that comes to their minds is the literal meaning of the words that make it up, that is, a research study that combines or mixes different methods. It is an informal meaning with which DIF researchers are quite familiar, given the long tradition in the field of (a) conducting DIF analysis with different DIF models looking for converging results, or (b) performing DIF studies by some kind of judgmental review together with DIF statistical methods.

The current agreed meaning of the MM research term conforms at first sight to that first impression, and has been informally used for years to refer to studies using different methodologies, mainly those that provide qualitative and quantitative data. However, if we refer explicitly to MR as a research paradigm or methodological framework, we need a normative definition and a set of core characteristic that help to distinguish what it is and is not.

Looking briefly at the history of MM research, almost every reference book on MM research refers back to the division that emerged in the twentieth century between researchers particularly in social sciences as a consequence of the quantitative-qualitative debate (Gorard & Symonds, 2010). Creswell and Plano Clark (2011) present a detailed table of the historical development of MM research with five periods: from the “formative period” in the 1960s with landmark writings like the article by Campbell and Fiske (1959) introducing the use of multiple quantitative methods in validation studies, through the “paradigm debate” and “procedural development” periods in the 1980s and 1990s (Bryman, 1988), and ending with an current overlap between the “expanding” and “reflective periods” roughly in the last two decades (e.g., Tashakkori & Teddlie, 2003).

In the last two decades, the history of the development of MM research has been shaped by a very active group of researchers committed to the third paradigm by means of books, papers, conferences, and what we can consider the official organ of the movement: the *Journal of Mixed Methods Research* (JMMR). For example, the

publication of the *Handbook of Mixed Methods in Social and Behavioral Research* (Tashakkori & Teddlie, 2003) is commonly considered by the research community as the legitimization of MR. Something similar happened to some of the MM editorials published in the JMMR. Some examples include the editorial by Creswell and Tashakkori (2007), in which they analyzed the different researchers' perspectives on how MM studies were performed, and a recent publication by Fetters and Freshwaters (2015) giving clear indications to authors on the characteristics of the manuscripts welcomed in the journal to be considered for publication.

The theoretical foundations of the MR paradigm has been and still is a topic for debate. Creswell and Plano Clark (2011) list and describe the philosophical foundations and, in their own words, worldviews behind the following paradigms: post-positivism, constructivism, participatory, and pragmatism. The last is frequently cited to support one of the main slogans of the research paradigm: that research aims dictate the choice of methods. Johnson and Christensen (2008) place pragmatism at the core of the MR paradigm since it states that what really matters is not whether researchers consider themselves as quantitative or qualitative but whether research methods used achieve the research objectives. Pragmatism allows research to avoid what MM research advocates consider the big limitation of the two previous paradigms: the methodological exclusivity.

The search for a normative definition of MM research has been also constant through the expanding period of the paradigm. Green, Caracelli, and Graham (1989) reviewed studies performed in different fields finding that researchers pay attention to some aspects of the MM research but not to other equally important aspects, and recognize the need for an agreement on requirements of a true MM research study. After a debate on the pro and cons of different definitions in this period, Creswell and Plano Clark (2007) proposed the definition currently that more adherents receive:

As a methodology, it involves philosophical assumptions that guide the direction of the collection and analysis and the mixture of qualitative and quantitative approaches in many phases of the research process. As a method, it focus on collecting, analyzing, and mixing both quantitative and qualitative data in a single study or series of studies. Its central premise is that the use of quantitative and qualitative approaches, in combination, provides a better understanding of research problems than either approach alone. (p. 5)

It is not difficult to find the two key elements of the definition: mixing through all research phases, and quantitative and qualitative data.

### ***Are DIF Researchers Ready to Perform MM DIF Validation Studies?***

Aware that the normative current definition should incorporate many different aspects and viewpoints, Creswell and Plano Clark (2011) rely on a list of core characteristics of true MM research. The list could work as a kind of check-list to consider if MM research comes in time to advance DIF research. Table 11.1 presents the core characteristics of MM research and our assessment for DIF studies.

**Table 11.1** Assessment of MM research core characteristic in DIF studies

Core characteristics of MM research		Assessment	Comment
A	Collect and analyze persuasively and rigorously both qualitative and quantitative data (based on research questions)	Done	DIF studies with judgmental reviews and DIF statistics are common
B	Integrates the two forms of data concurrently by combining them, sequentially by having one build on the other, or embedding one within the other	Pending task	There is a lack of designs that guide researchers when mixing qualitative and quantitative data
C	Gives priority to one or to both forms of data	Done	There have been a clear priority for quantitative data given the political and social contexts of most of the DIF studies
D	Uses these procedures in a single study or in multiple phases of a program of study	Done	There is a growing number of DIF studies that use different procedures in a single study
E	Frames these procedure within philosophical worldviews and theoretical lens	Pending task	New DIF models of the third generation convey the “theoretical lens” needed
F	Combines the procedures into specific research designs that direct the plan for conducting the study	Pending task	MM DIF validations need specific research design to reach a true integration

To sum up, DIF research is ready to advance within the methodological framework of MM research. From a methodological perspective, DIF researchers should work on the pending tasks to move the research area forward, but conducting studies in which integration is present from the formulation of the research problem, and from then plan and conduct their studies looking at integration as the key characteristic.

### Getting Integration into a MM DIF Validation Study

At this point it seems clear that the defining characteristics of MM research are: (a) the collection and analysis of both qualitative and quantitative data, and (b) the integration of findings and drawing inferences based on the qualitative, quantitative, and MM findings (Creswell, 2015). Therefore, the real challenge when conducting a MM study is integration. The rationale for integration is the first issue that researchers should address while planning a MM study. The second big issue is the set of decisions that have to be made in planning the research integration through all phases of the study. Our next task is to help DIF researchers in dealing with both issues. The arguments and contents we present are based on our own experience conducting MM DIF validation studies, and our reading of the normative practices in MM research.



### ***Rationale for a MM DIF Validation Study***

No matter how much we are convinced of the benefits for research on DIF that can come from the MM research framework, we also think that MM research is not the best methodological choice for all DIF studies. In fact, there are lots of DIF studies for which quantitative methods are and will be the best. Three of the five general purposes of third generation of DIF proposed by Zumbo (2007), can continue to be addressed with quantitative DIF methods: purpose 1, fairness and equity in testing, purpose 2, dealing with a possible threat to internal validity, and purpose 5, investigating the lack of invariance. In addition, DIF studies aimed at analyzing polytomous or dichotomous impact of DIF on test scores (e.g., Hidalgo, Benítez, Padilla, & Gómez-Benito, 2015), and those devoted to improve the detection of DIF and differential tests functioning, can be performed by quantitative methods.

The rationale for integration should be posed in terms of the potential advantages of combining both forms of data. Creswell and Plano Clark (2011) list several situations in which the need for integration is clear. Among such situations, there are two that readers will recognize has been mentioned in previous sections of the chapter: (a) one source of data may be insufficient to reach study aims, and (b) a need exists to explain initial results. Coming back to the purposes of the third generation of DIF (Zumbo, 2007), both situations are commonly recognized by researchers interested in purpose 3, investigating the comparability of translated and/or adapted measure understood as a matter of construct comparability, and purpose 4, trying to understand item response processes, when considering personal, contextual, cultural and social factors as potential sources of DIF.

DIF researchers interested in using DIF research to understand response processes in validation should not find difficulties in justifying their option for MM research if they put down the rationale in terms of the above needs and research purposes.

### ***Approaches to Integration in a MM DIF Validation Study***

Given that integration is the real challenge for MM research, considerable work has been done to guide researchers. In fact, MM design has been one of the most productive topics in the area. In this section, we briefly present one classification of MM design that have had a wide impact on the field, and then we focus on how to achieve integration in a MM DIF study resorting to one of the most recent proposed guidelines.

The longest-lasting classification of MM design was proposed by Creswell (1995). The classification is based on two dimensions, sequentiality and dominance. According to sequentiality, MM studies can be simultaneous or sequential, meaning qualitative and quantitative methods can be applied in parallel or stages. Dominance describes the priority, that is, one of the methods can be more dominant than the

other in terms of relevance for the research question, resources needed, etc., or both be equally relevant. In order to facilitate the communication, Creswell (1995) presented a systematic approach for representing MM design still currently used in published papers, articles, and reports. The MM design carried out in the study is represented by a combination of capital vs. lower case letters to represent dominance, plus symbols “/” or “+” to express sequentiality. For example, Maddox, Zumbo, Tay-Lim, and Qu (2015) combine ethnographic transcript and DIF analysis to examine how Mongolian respondents cope with three items of the United Nations Educational, Scientific and Cultural Organization (UNESCO) Literacy Assessment and Monitoring Programme (LAMP). If the authors would have wanted to classify their research design according to Creswell’s (1995) proposal, they may use the combination of letters and symbols “QUAN + QUAL” for a sequential MM design in which the quantitative DIF analysis and the observation were equally relevant, and the observations provide post-hoc explanation for DIF results.

Fetters, Curry, and Creswell (2013) update and propose a guide for achieving integration at the design, methods, interpretation, and reporting levels. Table 11.2 is an adaptation of that guide with DIF researchers in mind. We have selected the approaches at the different levels that are considered most promising to research on DIF causes at the current state-of art in the area. Of course, new innovative DIF methods could be supported by integration approaches not included in Table 11.2.

Starting with the integration at the design level, in an exploratory sequential DIF study, researchers should first collect and analyze qualitative data, then use the findings to inform quantitative data collection and analysis. Judgmental review to identify possible factors that could impact differently on item performance of subgroups of the target population could be planned following an exploratory sequential design.

DIF studies with an explanatory sequential design involve first collecting and analyzing quantitative data, and then using quantitative results to inform qualitative data collection and analysis. For example, the DIF research by Maddox et al. (2015) presented above can also serve as an illustration of an explanatory sequential design.

Integration at the method level can be fruitful for DIF research at connecting and building approaches. Connecting occurs when one makes a kind of data link to the other by the sampling frame. This is the most common case of MM design in cross-lingual or cultural survey or testing projects. Connecting leads to sampling for qualitative methods based on the quantitative data collection or analysis. Padilla, Benítez, and Castillo (2013) performed a MM evaluation project for the Spanish

**Table 11.2** Approaches to integrations for MM DIF validation studies

Integration level	Approaches for MM DIF validation studies
Design	Exploratory sequential
	Explanatory sequential
Methods	Connecting
	Building
Interpretation and reporting	Narrative

National Health Survey. Participants for cognitive interviewing qualitative method were recruited considering respondent profiles for which psychometric analysis revealed possible biases.

As another example, Zumbo et al. (2015) performed what they call a mixture DIF analysis. Within the ecological model of item responding they resort to latent class logistic regression to find out the explanatory power of different predictors of DIF in the English and French versions of the 2009 Program for International Student Assessment (PISA). Building occurred when authors used survey data from the Student Questionnaire for the cognitive processes, person characteristics, teacher, classroom and context factors, and ecology outside the school, as possible predictors of DIF. Data collection and analysis of Student Questionnaire was informed by the DIF analysis.

The most promising approach to integration at the interpretation and reporting level is narrative. Among the options for narrative integration, the so-called contiguous approach involves the interpretation and presentation of findings in a single reports, devoting different sections for the quantitative and the qualitative results. A detailed example of MM DIF validation study in which integration at the interpretation and reporting level has been reached by narrative is described in the next section.

The integration level included in the guide developed by Fetters, Curry and Creswell (2013) is the most general. Special attention is increasingly paid to integrative analysis of quantitative and qualitative data. Readers interested in deepening in this topic can follow the guidelines and description developed by Bazeley (2012).

## **An Example of a MM DIF Validation Study**

The aim of this section is to illustrate how to conduct a MM DIF validation study, focusing on response processes, following the methodological framework presented in the previous sections. We present a detailed description of a MM research study aimed at uncovering sources of DIF in the non-cognitive items of the Student Questionnaire for the 2006 PISA project, when comparing Spanish and US samples. The research carried out combines quantitative results from DIF analysis with qualitative findings from the cognitive interviewing (CI) method. The readers can find an introduction to CI methodology in a previous chapter in this book (Padilla & Leighton, 2017 *this book*).

The detailed presentation of the MM DIF study is organized keeping in mind two relevant and complementary objectives: (a) to provide a guide on how to implement the integration principle through all phases of the study, and (b) to follow best practices and indications on how to publish MM research (e.g., Fetters & Freshwaters, 2015). To achieve both objectives, we describe the study by Benítez and Padilla (2014). As many readers may have also experienced, the process of publishing the research informed us about how to communicate a MM research in a three-way very instructive process between reviewers, editor, and the authors.

### **Step 1: Providing a Rationale for a MM Study**

The dissatisfaction with the scant consolidation of results in the search for DIF causes was attributed to the inadequacy of the methodologies used and the complexity of the DIF phenomenon. What is more, the inability of traditional methods to provide information concerning the nature and location of DIF effects (Penfield, Alvarez, & Lee, 2009), could complete the rationale for the MM study.

The research aims for the study fit purpose 3 (investigate the comparability of adapted versions) and purpose 4 (trying to understand item response processes) of Zumbo's (2007) list of uses of DIF. We aimed to investigate the comparability of US English and Spanish versions of the some scales in the 2006 PISA Student Questionnaire. We intended to obtain evidence of the response processes and to link such evidence to findings about the locations of DIF effects provided by DIF analysis. In the end, we expected that the integration of a third generation DIF analysis and qualitative findings from cognitive interviewing provided solid evidence of the DIF causes.

### **Step 2: Integration at the Design Level**

The study followed an explanatory sequential design. First, we analyzed DIF using *Penfield's Differential Step Functioning* (DSF) framework (e.g., Penfield, 2010). DSF assumes a graded response model which uses a cumulative form, because in this model, the step function describes the probability that an examinee successfully advances to a score level equal to or greater than the chosen response. DSF allows researchers to analyze DIF at the item level and then identify the step or steps involved for the item flagged with DIF. Penfield et al. (2009) proposed a taxonomy with two dimensions, pervasiveness and consistency, to locate DIF effects and interpret DSF form.

To carry out the DIF analyses, data were obtained from the PISA database (OECD, 2006), in which responses of 17,405 participants from Spain and 4902 participants from the United States (US) were coded. We planned that quantitative results will inform qualitative data collection and analysis. In the second phase of the study, for the application of CI, 44 participants were recruited, 24 from Spain (15 women and 9 men) and 20 from the US (11 women and 9 men).

### **Step 3: Integration at the Method Level**

This MM DIF study can illustrate how to reach integration at the method level by using the connecting and building approaches at the same time. Integration by connecting was implemented performing sampling to mimic the characteristics of participants in the PISA study: students between 15 and 16 years who were in the final stages of compulsory education. In addition, CI participants in both countries were interviewed in their mother tongue (US English and Spanish spoken in Spain), and interviews took place in Chicago (US) and Granada (Spain).

Regarding the integration at the method level by building, we developed interviewing protocols considering DIF results obtained in the quantitative phase. To illustrate this point, we need to look at one of the Likert type response items in one of the scales included in the Student Questionnaire of the PISA 2006. Intended to capture "General value of science" construct, the item 1 stem is, "Advances in broad

science and technology usually improve people's living conditions," using responses *strongly agree*, *agree*, *disagree*, and *strongly disagree*. As DSF framework could provide insight into the locations of DIF effects (item stem or response options), we developed general and specific follow-up probes intended to capture the response processes to item 1. The following are examples of such probes for item 1:

- General probes: Let's start talking about how you answered the first questions. The first questions were about your opinion about science. They are questions about whether you have fun learning scientific issues, if you enjoy, etc.... what did you thought while answering these questions?
- Specific follow-up probe 1: What living conditions have you thought about when responding to the sentence, "Advances in broad science and technology usually improve people's living conditions?"
- Specific follow-up probe 2: To the questions, "Advances in broad science and technology usually improve people's living conditions," and "I am interested in learning about broad science," you have answered \_\_\_\_\_ (See and read the option marked by the participant in the statement). Tell me more about your answer, why did you answer that?

#### **Step 4: Integration at the Interpreting and Reporting Level**

As readers can see in Benítez and Padilla (2014), the integration at the interpretation and reporting level for this study used a narrative method. We presented results and findings integration following a contiguous approach, that is, quantitative and qualitative results were placed in different sections and then we integrated both. To illustrate this step, we use only DSF analysis results, and again item 1 for the general value of science construct.

In the quantitative subsection of the results, we presented and interpreted DSF results. In the case of item 1, DSF results showed a pervasive and constant DSF effect with a large magnitude against US respondents. What this DSF results pattern means is that US respondents need more ability, an attitude less favorable to science, to move from one response option to the next one across all alternative (DSF pervasive), and that the difference in the ability between both country groups is large, and more or less the same across all steps (DSF constant).

Qualitative findings provided by CI for item 1 are summarized in the qualitative subsection of the results. A sample of the transcripts of the responses to the specific follow-up probe 1:

- Spanish respondent (4): "Pues que cuando se hace algún invento si es útil nos va a facilitar la vida diaria, por ejemplo inventos como la bombilla o la televisión" ("When a new invention is done, daily life is easier... for example the light bulb or the television").
- US respondent (13): "Sickness or anything maybe they found a good advancement that this combination of drugs helps out more, or they have a bad advancement saying it doesn't work at all."

Similar themes and subthemes were identified across interviewees for each country group. In general, Spanish interviewees spoke about situations related with

daily life, mostly everyday objects, whereas US interviews social and health advances as results of sciences (e.g., new drugs, important medical treatments, means of transport, etc.).

While integrating DSF analysis results and qualitative findings, we came up with the idea that item 1 DIF could be explained by a difference in the meaning of the expression in the item stem "...people's living conditions." The DIF in item 1 could be associated with cultural factors that provoke US and Spanish respondent interpret different the expression in the item stem.

## Conclusions

There is a lot of interesting research work to do on DIF and item bias causes. The list of research purposes proposed by Zumbo (2007) is not only still valid but also a promising research agenda because of the new DIF models and validity theories. From our point of view, DIF and item bias research should not be pursued to only enforce testing fairness, but also to develop and achieve a deeper understanding of item responding processes, and hence as validity evidence. The more deep our knowledge of item responding processes, the more solid our interpretations and explanations of test score variation across test takers.

New methodological challenges are also on the horizon, such as new administration modes of tests and scales, web questionnaires, behaviors only developed in social networks, new communication devices, etc. All of these can change our traditional definition of item and testing conditions factors, and, as a consequence, DIF and item bias.

MM research can be a powerful methodological framework to integrate quite different kinds of data. Systematic integration, as opposed to haphazard integration, supports solid grounds for inference based on test scores, and provides substantial advances in measurement.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bazeley, P. (2012). Integrative analysis Strategies for mixed data sources. *American Behavioral Scientist*, 56(6), 814–828.
- Benítez, I., & Padilla, J. L. (2014). Analysis of non-equivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, 8(1), 52–68.
- Benítez, I., Padilla, J. L., Hidalgo, M. D., & Sireci, S. (2015). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education*, 29(1), 1–16.
- Bryman, A. (1988). *Quality and quantity in social research*. London: Routledge.

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.
- Creswell, J. W. (1995). *Research design: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.
- Creswell, J. W. (2015). *A concise introduction to mixed methods research*. Thousand Oaks, CA: Sage Publications.
- Creswell, J. W., & Plano Clark, W. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.
- Creswell, J. W., & Plano Clark, W. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.
- Creswell, J. W., & Tashakkori, A. (2007). Editorial: Developing publishable mixed methods manuscripts. *Journal of Mixed Methods Research*, *1*(2), 107111.
- Elosúa, P., & López-Jauregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing*, *7*(1), 39–52.
- Embretson, (Whitely) S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197.
- Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*(2), 24–35.
- Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113–148.
- Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs principles and practices. *Health Services Research*, *48*(6), 2134–2156.
- Fetters, M. D., & Freshwater, D. (2015). Publishing a methodological mixed methods research article. *Journal of Mixed Methods Research*, *9*(3), 203–213.
- Gorard, S., & Symonds, J. (2010). Death of mixed methods? Or the rebirth of research as a craft. *Evaluation and Research in Education*, *236*(2), 121–136.
- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274.
- Hambleton, R. K. (2006). Good practice for identifying differential item functioning. *Medical Care*, *44*(11), S182–S188.
- Hidalgo, M. D., Benítez, I., Padilla, J. L., & Gómez-Benito, J. (2015). How much polytomous item bias make total-group survey score comparisons invalid? *Sociological Methods and Research* 1–19.
- Hidalgo, M. D., & Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopaedia of education* (3rd ed., pp. 36–44). Oxford, UK: Academic Press.
- Johnson, B., & Christensen, L. (2008). *Educational research quantitative, qualitative, and mixed approaches*. Thousand Oaks, CA: Sage.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.
- Maddox, B., Zumbo, B. D., Tay-Lim, B., & Qu, D. (2015). An anthropologist among the psychometricians: Assessment events, ethnography, and differential item functioning in the Mongolian Gobi. *International Journal of Testing*, *15*(4), 291–309.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: MacMillan.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297–334.
- OECD (2006). Database - PISA 2006. Retrieved from <http://www.oecd.org/pisa/pisaproducts/databasepisa2006.htm>
- Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144.



- Padilla, J. L., Benitez, I., & Castillo, M. (2013). Obtaining validity evidence by cognitive interviewing to interpret psychometric results. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 9(3), 113–122.
- Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement*, 47(2), 129–149.
- Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education*, 22(1), 61–78.
- Sireci, S. G. (2012, April). “De-constructing” test validation. Paper presented at the annual conference of the National Council on Measurement in Education, Vancouver, BC.
- Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, 19(2–3), 170–187.
- Tashakkori, A., & Teddlie, C. (2003). The past and future of mixed methods research: From data triangulation to mixed model designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 671–701). Thousand Oaks, CA: Sage.
- van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. R. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology*. New York: Cambridge University Press.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223–233.
- 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, NC: IAP-Information Age Publishing, Inc..
- Zumbo, B. D. (2015, November). *Consequences, side effects and the ecology of testing: keys to considering assessment ‘In Vivo’*. Keynote address, the annual meeting of the Association for Educational Assessment – Europe (AEA-Europe), Glasgow, Scotland. [<https://youtu.be/OL6Lr2BzuSQ>]
- Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, 5(1), 1–23.
- Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera, O. L., & Tanvinder, K. A. (2015). A methodology for Zumbo’s third generation DIF analysis and the ecology of item responding. *Language Assessment Quarterly*, 12(1), 136–151.

# Chapter 12

## Cognitive Interviewing and Think Aloud Methods

José-Luis Padilla and Jacqueline P. Leighton

Although a case could be made that the need for explanations of item responses has been around since the origins of validity theory, the 1999 edition of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), can be considered the official birth certificate of validity evidence based on response processes as a source of validity evidence. Previous relevant references can be traced to a recommendation by Messick (1990) to look at how subjects cope with items and tasks to identify processes underlying item responses, efforts by Embretson (1983) linking cognitive psychology to item response theory, or even the earliest definitions of validity, if we just take an interest in knowing “what the test measures.”

At the same time that professionals, researchers, and testing organizations have been incorporating research on response processes into their test development and evaluation practices, the current *Standards* (AERA, APA, & NCME, 2014) maintains response processes among the five sources of validity evidence. On the downside, the latest *Standards* has not gone further in providing more indications than in the previous edition on how to obtain solid validity evidence based on response processes. Systematic reviews of validation studies reveal that few studies are conducted to obtain validity evidence using response processes. Cizek, Rosenberg, and Koons (2007) found that validity evidence based on participants’ response processes were studied only in 1.8% of the papers. Zumbo and Shear (Shear & Zumbo, 2014; Zumbo & Shear, 2011) showed a higher presence but still a minority compared with the other sources of validity evidence; for instance, in the medical

---

J.-L. Padilla (✉)  
University of Granada, 18071 Granada, Spain  
e-mail: [jpadilla@ugr.es](mailto:jpadilla@ugr.es)

J.P. Leighton  
Center for Research in Applied Measurement and Evaluation (CRAME), Department of Educational Psychology, Faculty of Education, University of Alberta,  
6-119D Education North Building, Edmonton, AB T6G 2G5, Canada  
e-mail: [jacqueline.leighton@ualberta.ca](mailto:jacqueline.leighton@ualberta.ca)

outcomes field, only 14% of the validation studies were aimed at obtaining evidence of the response processes.

The lack of experience, consolidation of best practices, and recommendations on how to obtain evidence of response processes can lead to missed opportunities provided by new conceptual and methodological developments in validity theories and validation methods. Among the various validation methods that can provide evidence of response processes, this chapter is devoted to cognitive interviewing (CI) and think aloud methods.

The target audience of this chapter is professionals and researchers looking for methodological guidance to perform validation studies by using CI and think aloud methods. In the chapter, we (a) describe the state-of-the-art in conducting think aloud and CI studies, (b) describe similarities and difference between the methods, and (c) demonstrate how both methods can provide validity evidence of response processes.

CI and think aloud methods are described in the context of educational and psychological testing. Both methods are often applied in survey research too, mainly as pre-testing methods to fix problems and improve survey questions. In fact, as we discuss in the following sections, both methods have common origins and not as distant developments as it might seem. We intend to provide arguments to distinguish between both methods to help researchers to make informed decisions about which method can be more useful considering the aims of the validation study.

We think that such validity evidence can be understood from a de-constructed view of validity (e.g., Kane, 2013; Sireci, 2012) to a more contextualized and pragmatic explanation validity framework (Stone & Zumbo, 2016; Zumbo, 2009). Throughout the chapter, we will also present studies to illustrate the content and how to apply think aloud and CI methods.

## **Introduction and State-of-the-Art in Conducting Cognitive Interviewing (CI)**

### ***CI History and Overview***

Before starting with a short history of the CI method, we should present a definition and a clear description of how the method is usually applied. The need for a definition is evident given that the term is also common in fields far from educational testing and psychological assessment, like law enforcement, where CI is a police resource to check witness reliability. What is more, CI emerged as a question evaluation method in the survey research field. Therefore, readers should pay attention to translating definitions of CI into the educational testing and psychological assessment context.

Although there is no universally accepted definition of CI, a wide consensus exists about what Beatty and Willis (2007) think CI involves: “the administration of

draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends” (p. 287). The first task for readers is to change ‘survey question’ to ‘test items’ or ‘scale items’. A couple of years later, Willis (2009) stated that CI “... is a psychologically-oriented method for empirically studying the way in which individuals mentally process and respond to survey questionnaires. Cognitive interviews can be conducted for the general purpose of enhancing our understanding of how respondents carry out the task of answering survey questions” (p. 106).

Highlighting the core elements in both definitions of CI allows us to recognize potential benefits from CI in validation studies of test score interpretations: (a) CI is a psychologically-oriented method for investigating respondents’ mental processes while answering test and scale items, (b) CI data can be useful for examining the quality of item responses, and (c) CI can help determine whether items are capturing the intended behaviors. The next section presents studies that illustrate these core elements.

Commonly, CI pre-testing evaluation studies in survey research consist of conducting in-depth interviews following an interview protocol with a small, purposive sample of 10–30 respondents. First, respondents answer the target survey questions; that is, the questions to be pre-tested, and then they respond to a series of follow-up probes that vary from general and open probes, like “*What were you thinking?*” or “*How did you come up with that?*” to much more scripted and specific follow-up probes, such as “*What does the term/word (...) mean to you?*” or “*How did you calculate (...)?*” Problems with the ‘question-and-answer’ process are usually identified and analyzed from the respondents’ narratives in the cognitive interviews.

As Miller (2014) points out, CI, by asking respondents to describe how and why they answered survey questions as they did, provides evidence not just to fix questions but also to find out the ways respondents interpret questions and apply them to their own lives, experiences, and perceptions. Miller’s (2014) interpretative view of CI methodology from survey research, coincides beyond expected with the broadest conceptions of validity theory in educational and psychological testing. For example, the contextualized and pragmatic explanation validity framework (Zumbo, 2009) expands opportunities for CI as a validation method to obtain evidence of response processes and to examine equivalence and sources of bias in cross-cultural research (Benítez & Padilla, 2014). However, we need to briefly summarize the evolution of CI before going into details of that proposal.

Almost all manuals and introductory articles on the CI method point out the Cognitive Aspects of Survey Methodology (CASM) conference (Jabine, Straf, Tanur, & Tourangeau, 1984) as a critical event in the history of CI. Presser et al. (2004) also identified the influential contribution of the Loftus’ (1984) post-conference analysis of how respondents answer questions about past events. Such an analysis relied on the think-aloud technique to studying the solving of problems developed by Ericsson and Simon (1980). So influential was Loftus’ (1984) work that, since then, the think-aloud technique has been closely linked to CI either as a

theoretical basis for the CI method (e.g., Willis, 2005), or as a data collection procedure along with verbal probing to conduct CI (e.g., Beatty & Willis, 2007).

After the CASM conference, and relying heavily on cognitive theory, cognitive laboratories devoted to testing and evaluating survey questions were established first at several U.S. federal agencies (e.g., the National Center for Health Statistics, the U.S. Census Bureau), and then at Statistics Canada and equivalent official statistics institutes like Statistics Netherlands, Statistics Sweden, etc. (Presser et al., 2004). As we discuss in the next section, the role of federal agencies and official statistics institutes can explain why CI methodology is still mainly seen as a pre-testing method aimed at fixing problems with questions to reduce response errors (Willis, 2005). Federal agencies and official statistics institutes have shaped CI methodology in survey research similarly to the way that testing companies have modeled research on item bias and differential item functioning (DIF).

Nowadays, CI practitioners and researchers live off of the advancements that the CASM conference brought to the study of measurement errors in survey research. The CASM movement sets the idea that respondents' thought processes must be understood to assess validity (Schwarz, 2007). Later, the inclusion of motivational elements to information-processing perspectives produced a major evolution. For example, Krosnick (1999) introduced the construct of "satisficing" to account for the tendency of most respondents to choose the first satisfactory or acceptable response option rather than options reflecting full cognitive effort. More comprehensive models of the question-and-answer process are on the way to take context, social, and cultural elements into account, support the rationale behind the method, and expand the range of validation research questions that could be addressed by CI (e.g., Shulruf, Hattie, & Dixon, 2008).

## *CI Approaches and Theories*

From the short introduction above to CI, it should be clear that CI is a qualitative method used to examine the question-and-answer process carried out by respondents when answering survey questions. Even though distinguishing between different purposes for conducting CI in survey research can be difficult, such a division can help us find out the ways in which CI can provide evidence of response processes for validation studies in testing and psychological assessment. Willis (2015) differentiates between two apparently contrasting objectives: reparative vs. descriptive cognitive interviews. With slight changes in the labels, the distinction can be easily found in the literature when the purpose of CI is under debate (e.g., Chepp & Gray, 2014; Miller, 2011). The reparative approach corresponds to the original need for identifying problems in survey questions and repairing them. Traditionally in cognitive laboratories and official statistics institutes, it has been the practice of CI projects to answer, let us say, quantitative questions within a qualitative method: "How many problems does the target question have?" or "Which percentage of CI participants reveal such problem?" In contrast, the descriptive approach represents

CI projects whose aims are to find out how respondents mentally process and answer survey questions instead of just uncovering response errors. Advocates of this approach argue that CI should be planned to discover what a survey question is truly capturing, that is, how survey questions function as measure of a particular construct (e.g., Chepp & Gray, 2014; Ridolfo & Shoua-Glusberg, 2011).

The descriptive approach is in line with our proposal to rely on the CI method to obtain validity evidence related to response processes associated with test and scale items. There is a solid argument for the parallelism between the more comprehensive objective of discovering what the survey question is truly capturing and the 2014 *Standards* definition for validity evidence of response processes: “Some construct interpretations involve more or less explicit assumptions about the cognitive process engaged in by test takers. Theoretical and empirical analysis 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” (p. 15). The following indication to questioning test takers about their performance strategies or responses to particular items opens the door to applying CI methodology from a descriptive approach to obtain validity evidence of response processes.

The question now is if there is a theory to support CI methodology. Willis (2015) proposes to distinguish between what he calls a theory of the phenomenon, that is, how people respond to survey questions, and a theory of the method, a theory that supports the use of CI to test and investigate survey response processes. Starting with the theory of the phenomenon, the CASM view, such as it is exposed by the four-stage cognitive model by Tourangeau (1984), has been and still is the most cited cognitive theoretical framework of response processes to survey questions. The model presents a linear sequence from when the survey questions are presented to the respondent to the selection of a response: (a) comprehension of the question, (b) retrieval of relevant information, (c) judgment/estimation processes, and (d) response. More recently, elements of disciplines like linguistics, anthropology, or sociology have been incorporated to account for the effects of context, social, and cultural factors, etc., on response processes (e.g., Chepp & Gray, 2014).

Regarding the theory of the method, Willis (2015) thinks that CI still relies on Ericsson and Simon’s (1980) defense of think-aloud interviews to obtain access to the functioning of cognitive processes. For Willis (2015), the idea that persons who spontaneously verbalize their thoughts provides a ‘window into the mind’ remains as the theoretical base for CI, what blurs borders between think aloud and CI methods, and explains why the CI method is sometimes referred as ‘think-aloud interviews’. Due to the lack of empirical evidence of the veracity of verbal reports provided by CI, current contributions from other social science disciplines (e.g., ethnography, sociology), and the growing application of CI cross-cultural research, CI is starting to be viewed as a qualitative method and something more than just ‘cognitive’ (e.g., Willis & Miller, 2011).

Among the qualitative approaches to CI, one of the most promising is the interpretive perspective within the framework of Grounded Theory (e.g., Ridolfo & Shoua-Glusberg, 2011). The rationale behind of this approach is the production of a full

range of themes in CI data and the need to study the CI topic (in this case, the response processes to a survey question) until saturation is reached. Briefly, from an interpretative perspective, the topic is what the meaning is for the respondent, and meaning is socially constructed by the respondent in a particular moment and a particular social location. A detailed treatment of the interpretative perspective, in the context of CI, can be found in Chepp and Gray (2014). Miller, Willson, Chepp, and Padilla (2014) present an exhaustive description of the main phases and aspects of CI methodology from an interpretative perspective. The next section of the chapter presents examples of studies conducting CI as a validation method in the context of educational testing and psychological assessment from an interpretative perspective.

### ***Conducting a CI Validation Study: Main Phases, Procedural Issues, and Examples***

To help researchers in making informed decisions on the adequate method—either CI or think-aloud—to obtain validity evidence of response processes, this section presents the main phases and some practical issues on CI. Fortunately, readers interested in all procedural details of the CI method can be referred to a set of books published in the last years: Collins (2015), Miller, Chepp, Wilson, and Padilla (2014), Willis (2005), and Willis (2015). From different approaches to CI, these books were mainly written for an audience of survey researchers. Considering the aim of the chapter and our experience, we have selected and adapted the contents that can be more useful for a validation study in educational and psychological testing. The three main phases to be considered when planning a CI validation study are discussed in the following.

**Fitting the CI Study into the Overall Validation Project** The introduction to the evidence based on response processes in the validity chapter of the latest *Standards* (AERA et al., 2014), include indications that can help in responding to the question, when should a validation study based on response processes be conducted? Obviously, the question is relevant for both CI and think-aloud methods, and should be responded to before considering applying any of them. The indications point out the validity research questions for which both methods can be appropriate: “Evidence of response processes can contribute to answering questions about difference in meaning or interpretations of test scores across relevant subgroups of test takers” (AERA et al., 2014, p. 15).

Benitez and Padilla (2014) propose three general propositions which could be examined by evidence provided by CI: (a) the performance of test takers reflects the psychological processes delineated in test or scale specifications, (b) the processes of judges or observers while evaluating test taker’s products are consistent with the intended interpretations of scores, and (c) relevant subpopulations of test takers defined by demographic, linguistic, or cultural groups do not differ in the response processes to test and scale items.



CI fits into different moments of overall validation projects and can be integrated with other validation methods. To support test uses or propositions involved in a validity argument can require multiple strands of quantitative or qualitative validity evidence. For example, Castillo and Padilla (2013) conducted CI to interpret differences in the factor structure of a psychological scale intended to measure the construct of family support. Therefore, the integration of different validation methods, among them the CI method, should be addressed in a systematic way from the beginning of the validation project. A mixed methods research framework, introduced by Padilla and Benitez (in this book), offers a path to reaching such integration.

**Planning CI** To contrast with CI practice in survey research, in which single survey questions are the “target,” we intend to obtain evidence of response processes of multi-items tests or scales. Of course, researchers can focus on particular items, but test takers respond to tests and scales as a whole. Conrad and Blair (2009) stated conditions in order that CI can provide evidence of a non-automatic processing of item scales. Unsurprisingly, to sum up such conditions, test takers should be aware of response processes and able to communicate about them during the interview. Planning CI involves taking care of many procedural issues. Next, we address the most important aspect of planning in the context of educational and psychological testing.

*Developing the Interview Protocol* A movie script can come to the reader’s mind as an example of an interview ‘protocol’. At the end of the day, a CI is an interview with two main characters: interviewer and respondent. To some extent, the comparison conveys the key role of the interview protocol. It consists of the introduction of the study to the respondents (e.g., statements of the research aims, main topics, responsible organization, confidentiality), information of the expected role for the respondent, and the probes. However, as a validation method, the interview protocol is much more than a script. The content of the interview protocol, structure, even its length, reflects the researcher’s approach to the CI method. To opt for a reparative vs. a descriptive approach to CI leads to very different interview protocols. A CI study from an interpretative approach develops an interview protocol which allow researchers to obtain the socially constructed meaning of the items for the respondent whereas, from a solving-problem perspective, the protocol intends to facilitate the evaluation of questions task. Table 12.1 outlines the bi-directional conditioning effects between the roles of the respondents and interviewers, and the kind of probes mostly included in the interview protocol.

Willson and Miller (2014) presented what we can call two oppositions to characterize the expected role of the respondents and the interviewers that condition the kind of probes included in the interview protocol. The respondents act as ‘evaluators’ when they are asked to evaluate parts of the question: stem, response options, or their own cognitive processes, while acting as ‘storytellers’, where “they are asked to generate a narrative that depicts ‘why they answered the question in the way that they did’” (Willson & Miller, p. 26). The second opposition sets a parallelism with the expected role of the interviewer as a ‘data collector’ or as a ‘researcher’.

**Table 12.1** Relation between interviewer and respondent

Role Respondent	Probes	Role Interviewer
Evaluator	Standardized and structured	Data collector
Storyteller	Spontaneous	Researcher

**Table 12.2** Examples of probes used in interview protocols

GENERAL PROBE:
P.1. Let's start talking about how you answered the first questions. The first questions were about how important aspects like "work," "family," "friends," etc., are for you. How did you answer these questions? What did you take into account for responding?
SPECIFIC PROBES:
P.2. One of the aspects was "family," what did come to your mind while responding? What persons did you think of?
P.3. Other aspect was "friends and acquaintances," you have answered _____ (See and read the alternative selected by the participant in statement), tell me more about your answer, why did you answer that.

If the interviewer is instructed to ask the same probes in the same way to every respondent, we have data collectors that do their best to avoid interviewer biases and preserve CI data accuracy. In contrast, the interviewer is a qualitative researcher when they "assess the information that he or she is collecting and examine the emerging information to identify any gaps, contradictions, or incongruences in the respondent's narrative" (Willson & Miller, 2014, p. 30). In this case, the interview protocol is open to what Willis (2005) called spontaneous or free-form probes to help interviewers in leading the interviewing.

Benitez, He, van Vijver, and Padilla (2016) conducted a CI study to obtain validity evidence of the response processes to some quality-of-life questions and scale items used in international studies, comparing Spanish and Dutch respondents. Table 12.2 presents a sample of the interview protocol for questions intended to capture how important aspects like family, work, friends, etc., are for participants. The sample includes a general probe and two specific probes. Interviewers were instructed to resort to the specific follow-up probes when interviewees' comments did not provide a full narrative of what items meant for them and how they had constructed their responses.

The books by Willis (2005) and Miller, Willson, et al. (2014) provide detailed descriptions of the different kind of probes, and how they determine not just interviewer and respondent roles, but also, as could not be otherwise, the CI data analysis.

*Recruitment* How many interviews and who should be the respondents are permanent concerns when researchers decide to conduct a CI validation study. Researchers should not forget that CI is a qualitative method. Thus, sampling is not a primarily numerical matter, but a purposive one. Learning from the survey research field, we can base sampling on demographic diversity or the topic covered by the items. For

example, if we want to obtain validity evidence of the response processes to a quality of life scale for people with disability, CI sampling should include people with different disabilities. The AERA et al. (2014) *Standards* reiterates the idea of comparing response processes “about difference in meaning or interpretation of test scores across relevant subgroups of test takers” (p. 15).

As a qualitative validation method, CI can benefit from criteria to respond to the question of sample size and composition: theoretical saturation and theoretical relevance (Willson & Miller, 2014). In the context of an educational testing or psychological assessment validation study, theoretical saturation implies that one keeps interviewing until research reaches a full understanding of how and why respondents answer the items and find potential difference across groups of respondents. With respect to theoretical relevance, along with respondents belonging to the relevant groups defined in the validity intended interpretations, researchers should consider including participants that can provide as much diversity as possible regarding response processes to the test items. As it is hard to avoid giving a number, in our experience, both criteria can be met with between 20 and 50 interviews.

*Interviewer Training* There is no simple answer to what competencies a good interviewer should have. Obviously, there are technical abilities and interpersonal skills that can make a difference when conducting the CI method. Willis (2005) described the technical background that can be helpful: (a) some type of social science education, (b) knowledge and experience in questionnaire design, (c) some exposure to the subject matter of the questionnaire, and (d) experience in conducting CI. The more experience interviewers have, the more capable they will be to manage and lead interviews. Willis (2005) also paid attention to the non-technical skills the interviewers should have: “the ability to be flexible, spontaneous, and cool under duress” (p. 130).

**Analyzing CI data** All major manuals of the CI method include a chapter devoted to CI data analysis. As readers may guess, different approaches to CI correspond with different types of analytic processes. Willis (2015) published a state-of-art book titled *Analysis of the Cognitive Interview in Questionnaire Design*, where the different analytical strategies, models, and critical issues in current analytic practices can be found.

We summarize here the analytic process from an interpretative approach to CI as a method to obtain validity evidence of response processes. Miller, Willson, et al. (2014) outline five incremental steps by which the reduction and synthesis process of CI data can be conceptualized: (1) conducting interviews to produce the interview text; (2) synthesizing interview text into summaries; (3) comparing summaries across responding to produce a thematic schema; (4) comparing identified themes across subgroups to produce an advanced schema; (5) making conclusions to produce a final study conclusion. From this perspective, analysis starts with the interview itself given that the interviewer acting as a researcher make analytic decisions along the way: identifying contradictions, following up respondent first responses, etc. Lastly, the main steps described follow an iterative process in practice: analysts go forward and backward through the analytic process (Miller, Willson, et al. 2014).

Benitez et al. (2016) followed the interpretative approach to analyze CI data obtained to compare response processes to quality-of-life questions and scale items between Spanish and Dutch respondents. For example, the researchers found a different interpretation pattern of the family concept. In contrast to Dutch participants, Spanish participants include within the family concept not just the immediate family, but also relatives and friends.

## **Introduction and State-of-the-Art in Conducting Think Aloud Interviews**

### *History and Overview of Think Aloud Interviews*

The think aloud interview is a psychological method used to collect data about human information processing, namely, problem solving. Problem solving has been defined as the goal-driven process of finding a solution to a complex state of affairs (Newell & Simon, 1972). Problem solving requires the manipulation of information to create something new and, therefore, is normally involved in higher-level skills found in Bloom's taxonomy (Bloom, Engerhart, Furst, Hill, & Krathwohl, 1956). The think-aloud interview can be a useful tool in determining whether test items or tasks elicit problem-solving processes. The think-aloud interview technique needs to be distinguished from cognitive labs, which are used to measure a wider array of response processes, especially comprehension (see Leighton, 2017a, 2017b). Cognitive labs are not the focus of this section and not discussed further.

The think-aloud interview has historical roots in experimental self-observation, a method used by Wilhelm Wundt (1832–1920) to systematically document the mental experiences of trained human participants to a variety of sensory stimuli. Unlike introspection, experimental self-observation was standardized to provide a structured account of the unobservable but systematic human mental experience. However, beginning in the 1920s, behaviorism became the dominant paradigm for studying psychological phenomena and only observable behavior was viewed as worthy of measurement. In the 1950s, the cognitive revolution, instigated by scholars such as Noam Chomsky and psychologists such as George Miller, Allan Newell, Jean Piaget, and Herbert Simon effectively replaced behaviorism as the dominant paradigm and methods to scientifically study mental experiences as accounts for human behavior and became a focus of interest (Leahey, 1992).

The think aloud interview as it is currently conceived was developed by two cognitive scientists, K. Anders Ericsson and Herbert Simon. In 1993, Ericsson and Simon wrote their seminal book *Protocol Analysis: Verbal Reports as Data* based upon a decade of their own research into the scientific study of human mental processing (e.g., Ericsson & Simon, 1980) and a review of previous research that was focused on the study of human mental processing. The 1993 book continues to be the major reference in the field. A careful reading of their book makes the following

unequivocal – inferences or claims about human problem solving are supported by data collected from think aloud interviews *only* when the interviews are conducted in a highly structured and systematic manner. In particular, the following conditions must hold: (a) the content of the interview must involve a problem-solving task, (b) the problem-solving task must require what is called controlled processing (i.e., processing that is not automatic but, rather in the participant’s awareness and open to verbalization) for its solution, and (c) the interview probes must be minimal and non-directive, without requests for elaboration and explanation, to allow the participant to verbalize concurrently. These three conditions must be met if the objective of the think-aloud interview is to collect evidence about human problem solving. If these conditions do not hold, claims or inferences about human problem solving are suspect at best and unwarranted at worst (Ericsson & Simon, 1993; Fox, Ericsson, & Best, 2011; Leighton, 2004). Hence, in the validity arguments created to bolster claims about test items measuring problem solving processes (e.g., in mathematical or scientific domains), the data from think aloud interviews can only serve as evidence of such claims if the data have been collected according to specific procedures, as discussed next.

### *Interview Sessions for Conducting Think-Aloud*

There are normally two sessions or parts to include in the think-aloud interview– the concurrent session and the retrospective session. Both involve unique interview probes. The details of these have been elaborated at length in past publications (e.g., Ericsson & Simon, 1993; see also Leighton, 2004, 2013, 2017b for instructions), but a summary bears repeating here. First, the concurrent session of the interview is most important and characterized by requesting the participant (or examinee) to verbalize his or her thoughts aloud in response to a problem-solving task. The objective is to have the participant (or examinee) solve the task and simultaneously verbalize the mental processes being used, in real time, to solve it. During this part of the interview, the interviewer should not interrupt with any questions (e.g., *Can you elaborate on why you are drawing a diagram to solve the problem?*) that would disrupt the flow of problem solving and thus verbalization or lead the participant to consider a distinct problem solving route (e.g., *Why not consider a diagram in solving the problem?*) not previously contemplated. The only probes the interviewer should use during this session are non-directed reminders to the examinee to verbalize thoughts as he or she is solving problem. For example, permissible non-directive probes would include, “*Please keep talking*” or “*Please remember to verbalize.*” The interviewer should avoid directive probes such as “*What are you thinking*” because this probe is a question that takes focus away from the task and requires the examinee to respond to the interviewer. If these protocol or procedural details seem overly specified, it is deliberate. True-to-life problem-solving processes are not necessarily robust to measurement–meaning that they are difficult to measure accurately. This is because these processes take place in working memory and the

contents of working memory are fleeting (see Ericsson & Simon, 1993). The data produced from this concurrent phase comprise a verbal report.

The second part of the think aloud interview is the retrospective session, and it is secondary in importance. It is characterized by having the examinee recount how he or she solved the problem-solving task. The retrospective session follows directly after the concurrent session and is initiated by having the interviewer request for the examinee to “*Please tell me how you remember solving the task.*” During the session, the interviewer may ask for elaboration and explanation of how the examinee remembers solving the task (e.g., *Why did you decide to draw the diagram?*). These elaborative questions are designed to help contextualize the verbal report the examinee provided during the concurrent session. The verbalizations an examinee provides during the retrospective session *are not* considered to be the primary evidence for supporting claims about problem solving (see Ericsson & Simon, 1993). This is because the retrospective session relies heavily on an examinee’s memory and does not capture the problem-solving process *in vivo*. One of the main weaknesses of verbal reports as evidence of problem-solving processes is the failure to follow protocol, namely, properly collect the reports during the concurrent session of the interview (see Fox et al., 2011; Leighton, 2004, 2013; Wilson, 1994). These failures will undermine the utility of verbal reports in validity arguments.

### ***Conducting a Think-Aloud Validation Study: Main Phases, Procedural Issues and Examples***

There are five phases for conducting think aloud interviews. The phases include: (1) cognitive model development; (2) instructions; (3) data collection using concurrent and retrospective probes; (4) coding of verbal reports using protocol analysis; and (5) generating inferences about participants’ response processes based on the data. Each of these phases involves specific methods or procedures. It is beyond the scope of the chapter to delve into these details, but interested readers are referred to Leighton (2017b) for a fuller exposition. At this point, it is important to repeat that the phases of the think-aloud interview differ from those used in ‘cognitive labs’, a variant interview of the think-aloud method that is used to measure comprehension rather than problem solving (the reader is again referred to Leighton, 2017b for a full exposition on the differences between think aloud interviews and cognitive labs). In this section, the main phases of the think aloud are summarized with brief presentation of procedural issues, with examples.

**Cognitive Model Development** Think-aloud interviews can yield a significant amount of verbal report data to analyze. Often researchers can become overwhelmed with the extent of the report data and what to focus on and evaluate as evidence of response processes. This is one reason why the first step in conducting a think-aloud is to develop a cognitive model, or some type of flowchart that outlines the knowledge and skills expected to underlie performance. The cognitive model does not

have to be complicated. However, it should illustrate the response processing expected as it will serve as a roadmap for identifying the knowledge and skills of interest in the verbal reports. If the model fails to fully or partially illustrate what is observed in the reports, then the model is refined based on the data. Leighton, Cui, and Cor (2009) provide an example of an expert-based cognitive model from expert analysis. It is a coarse-grained model developed by an expert for 15 algebra multiple choice SAT items; finer-grained models can be developed but can present challenges for inter-rater reliability. The cognitive model is the first step in structuring the measurement of response processes.

**Think-Aloud Instructions** Think aloud interviews, as originally conceived by Ericsson and Simon (1993), are used primarily to measure problem solving processes. The instructions used to initiate the interview must therefore be administered to ensure (a) participant comfort with verbalizing problem solving processes (a practice phase), (b) the minimization of participant response bias (indicate non-evaluation), and (c) participant focus on concurrent verbalization (concurrent probes). Because participants can easily become self-conscious about problem solving in front of an interviewer, it is suggested that participants be given time to practice projecting their voice. Often, participants will express and show comfort verbalizing with practice tasks, but when they begin the actual task of interest, will go silent. This often occurs as simultaneously thinking through the task information and verbalizing burdens working memory resources. However, participants need to be reminded to verbalize as they think through the task as this is the target of what is being measured, even if this means slowing down how they solve the task.

**Data Collection Using Concurrent and Retrospective Probes** As mentioned previously, there are two parts to the think-aloud interview—a concurrent session and a retrospective session. Each session has unique probes to ensure that the target response processing, namely problem solving, is being measured as intended. As explained in Leighton (2017b), only minimal, non-obtrusive and non-directed probes are permissible in the concurrent session, where the actual problem solving of interest is being observed and measured. Permissible concurrent probes include “*Please keep talking*” and “*Remember to continue talking.*” Elaborative probes that involve “why” or “how” questions are not permissible as they are often directive, obtrusive, and may bias the problem solving in which the participants is engaging. For example, probes such as “*Why did you do this?*” or “*How did you decide to select this option?*” can function as a source of feedback and influence the direction of problem solving. Elaborative probes are permissible during the retrospective session given that this session is designed to provide complementary but secondary evidence in relation to the problem-solving response processing (see Leighton, 2017b).

**Coding of Verbal Reports from Think Aloud Interviews** When verbal reports are collected as part of a validation project, the integrity of the data alongside interpretations or observations made from the data must be carefully considered and verified. For this reason, the coding of verbal reports should follow a rigorous and



standardized process that includes multiple raters and computation of inter-rater reliability (see Leighton, 2017b for details). First, a coding manual needs to be created based on the cognitive model developed for the task of interest. Second, the coding manual should include the set of knowledge and skills expected and examples of types of verbalizations that would present as evidence of these knowledge and skills. Third, at least two raters need to be independently trained to use the manual to categorize a proportion of the verbalizations of interest (e.g., 15–25%). Fourth, the raters need to be naïve to the objectives of the think aloud interviews, in terms of task difficulty, discrimination, potential differential item functioning, etc. Fifth, the initial agreement between the raters needs to be computed and, if low, further training undertaken to increase reliability of verbal report interpretation. Sixth, once inter-rater agreement is acceptable (e.g., Kappa of .60 or greater; see Landis & Koch, 1977), one rater can proceed to code the remainder of the verbalizations in the reports.

**Generating Inferences About Response Processes Related to Problem Solving** As already noted, when verbal reports become part of the validity argument for claiming that test-takers are engaging in specific problem solving processes, the integrity of verbal report data and observations made about the data must be verified. Claims made about problem solving processes cannot be made using any type of interview method (see Ericsson & Simon, 1993; Leighton, 2017a; Wilson, 1994). Although the think-aloud interview provides a tool for measuring problem solving, procedural deviations can undermine the data collected and claims (see Leighton, 2017b for cognitive labs and the target of measurement). Thus, generating inferences about response processes requires not only collecting verbal report data but also demonstrating that the procedures used to collect and interpret those data minimize bias, and are not subject to alternative interpretations and idiosyncratic conclusions. These issues are elaborated in Leighton (2017b).

## **Conclusion: Strengths and Weaknesses of Verbal Reports for Validity Arguments**

As indicated previously, the *Standards* (AERA et al., 2014) maintain the need to include evidence of response processes when generating validity arguments to support claims about skills, competencies, attitudes, beliefs, etc. that are difficult to observe or measure directly. While the *Standards* emphasize the need for evidence of response processes, the *Standards* do not describe how this evidence should be gathered and the best practices for gathering this evidence. Clearly, it can be assumed that evidence used to validate claims needs to be sound. The good news is that there is a solid base of past research on the conditions for gathering this evidence using different interview methods—cognitive interviews and think-aloud to name the two to which this chapter is devoted—and a growing body of research specifically in the domain of educational testing and increasingly so in psychological assessment, cross-cultural testing, etc.

A fundamental step in including interviews in validation arguments is to be clear about what type of response processing is being measured using the interview. Toward this end, at least for think-aloud interviews, it is necessary to identify the expected knowledge and skills expected *before* planning the interviews and administering tasks or items to examinees. Next, the integrity of the data has to be verified. It bears repeating that the strength of verbal report evidence is contingent on how the data were collected and how the data are interpreted. Coding manuals, raters, and checks on rater evaluations are key.

We expected that readers have obtained a clear picture of the main characteristics of CI and think-aloud methods. Even though origins and theoretical bases are closer than expected—both methods rely on the foundations developed by Ericsson and Simon (1980)—the evolution and current practices as it is exposed in this chapter, allow delimitating both methods. For example, while in think-aloud, the interview probes must be minimal and non-intrusive, CI is in general much more direct and intrusive by requesting elaborations from respondents, encouraging interviewing to look for contradictions in respondents' narratives, etc. Furthermore, the think-aloud interview is focused on the human problem solving domain, whereas CI not just comes from survey research but its use is growing in psychological assessments, cross-cultural research, etc.

Within any validity theoretical framework, researchers should be aware that the most important decisions to be made before collecting verbal reports using the think aloud interview is determining what are the response processes that test items are expected to measure and what are the appropriate procedures for collecting this evidence without biasing the data and the subsequent inferences. In contrast, CI could be conducted from, let us say, a more exploratory perspective to uncover what questions and scales are capturing.

Verbal report data, regardless of the method used to obtain them, are no different than any other data; quality rests with the methods used to minimize bias and avoid idiosyncrasies in interpretation. Both methods require that data are collected according to specific procedures. The difference is how think-aloud and CI understand and deal with such "bias." While think-aloud rely on rigorous and standardized inter-raters agreement evaluation, CI, as a qualitative method, trusts that transparency establishes credibility and validity through all phases. CI researchers should document any decision made while conducting the method, especially during CI data analysis, in order to achieve transparency.

## 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). *The standards for educational and psychological testing*. Washington DC: Author.

- Beatty, P., & Willis, G. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287–311.
- Benítez, I., He, J., van de Vijver, F. J. R., & Padilla, J. L. (2016). Linking extreme response styles to response processes: A cross-cultural mixed methods approach. *International Journal of Psychology*, 51, 464–473.
- Benítez, I., & Padilla, J. L. (2014). Analysis of non-equivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, 8, 52–68.
- Bloom, B. S., Engerhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York, NY: David McKay.
- Castillo, M., & Padilla, J. L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research*, 114, 963–975.
- Chepp, V., & Gray, C. (2014). Foundations and new directions. In K. Miller, S. Willson, V. Chepp, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 7–14). Hoboken, NJ: John Wiley & Sons.
- Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2007). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68, 397–412.
- Collins, D. (2015). *Cognitive interviewing practice*. London, UK: Sage.
- Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly*, 73, 32–55.
- Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–197.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215–251.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.
- Fox, M. C., Ericsson, A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137, 316–344.
- Kane, M. T. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, 50, 115–122.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50, 537–567.
- Jabine, T., Straf, M., Tanur, J., & Tourangeau, R. (Eds.). (1984). *Cognitive aspects of survey design: Building a bridge between disciplines*. Washington, DC: National Academy Press.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Leahey, T. H. (1992). *A history of psychology* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6–15.
- Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education*, 26, 136–157.
- Leighton, J. P. (2017a). Collecting, analyzing and interpreting verbal response process data. In K. Ercikan & J. Pellegrino (Eds.), *Validation of score meaning in the next generation of assessments*. Routledge.
- Leighton, J. P. (2017b). *Using think aloud interviews and cognitive labs in educational research*. Oxford, UK: Oxford University Press.
- Leighton, J. P., Cui, Y., & Cor, M. K. (2009). Testing expert-based and student-based cognitive models: An application of the attribute hierarchy method and hierarchical consistency index. *Applied Measurement in Education*, 22, 229–254.

- Loftus, E. (1984). Protocol analysis of response to survey recall questions. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 61–64). Washington, DC: National Academy Press.
- Messick, S. (1990). *Validity of test interpretation and use, Research report No. 90-11*. Princeton, NJ: Education Testing Service.
- Miller, K. (2011). Cognitive interviewing. In K. Miller, J. Madans, A. Maitland, & G. Willis (Eds.), *Question evaluation methods: Contributing to the science of data quality* (pp. 51–75). New York, NY: Wiley.
- Miller, K. (2014). Introduction. In K. Miller, S. Willson, V. Chepp, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 1–6). New York, NY: Wiley.
- Miller, K., Chepp, V., Willson, S., & Padilla, J. L. (Eds.). (2014). *Cognitive interviewing methodology*. New York, NY: Wiley.
- Miller, K., Willson, S., Chepp, V., & Ryan, J. M. (2014). Analyses. In K. Miller, S. Willson, V. Chepp, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 35–50). New York, NY: Wiley.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*, 136–144.
- Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly*, *68*, 109–130.
- Ridolfo, H., & Schoua-Glusberg, A. (2011). Analyzing cognitive interview data using the constant comparative method of analysis to understand cross-cultural patterns in survey data. *Field Methods*, *23*, 420–438.
- Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, *21*, 277–287.
- 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.
- Shulruf, B., Hattie, J., & Dixon, R. (2008). Factors affecting responses to Likert type questionnaires: Introduction of the ImpExp, a new comprehensive model. *Social Psychology of Education*, *11*, 59–78.
- Sireci, S. G. (2012, April). “De-constructing” test validation. Paper presented at the annual conference of the National Council on Measurement in Education as part of the symposium “Beyond Consensus: The Changing Face of Validity” (P. Newton, Chair), Vancouver, BC.
- Stone, J., & Zumbo, B. D. (2016). Validity as a pragmatist project: A global concern with local application. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice* (pp. 555–573). Newcastle, UK: Cambridge Scholars.
- 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.
- Willis, G. B. (2005). *Cognitive interviewing*. Thousand Oaks, CA: Sage.
- Willis, G. B. (2009). Cognitive interviewing. In P. Lavrakas (Ed.), *Encyclopedia of survey research methods* (Vol. 2, pp. 106–109). Thousand Oaks, CA: SAGE.
- Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. New York, NY: Oxford University Press.
- Willis, G., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods*, *23*, 331–341.
- Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science*, *5*, 249–252.

- Willson, S., & Miller, K. (2014). Data collection. In K. Miller, S. Willson, V. Chepp, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 15–33). New York, NY: Wiley.
- 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* (pp. 65–83). Charlotte, NC: Information Age Publishing, Inc.
- Zumbo, B. D., & Shear, B. R. (2011). The concept of validity and some novel validation methods. In *Northeastern Educational Research Association* (p. 56). Rocky Hill, CT.

# Chapter 13

## Some Thoughts on Gathering Response Processes Validity Evidence in the Context of Online Measurement and the Digital Revolution

Lara B. Russell and Anita M. Hubley

### The Digital Revolution and the Rise of Technology

The Digital Revolution refers to the rapid growth of both information and communication technologies as well as innovations in digital systems that have fundamentally changed and revolutionized the way people think, behave, communicate, share information, and work (e.g., <https://www.techopedia.com/definition/23371/digital-revolution>; Isaacson, 2014; Ramasubramanian, 2010). Internet use has increased dramatically over the past 15 years, from 400 million users in 2000 to an estimated 3.2 billion users in 2015 (International Telecommunications Union, 2015). Moreover, there is now greater variety in the kinds of devices used to access the internet, beyond desktop and laptop computers to devices such as tablets, netbooks, and mobile phones (e.g., Callegaro, 2010). The growth in mobile (cellular) phone use has also been dramatic; between 2000 and 2015, the number of cellular subscriptions worldwide rose from about 10 per 100 individuals to close to 97 per 100 individuals. Mobile broadband (wireless internet) subscriptions have also risen sharply, from under 5 per 100 individuals in 2007 to an estimated 47 per 100 individuals in 2015 (International Telecommunications Union, 2015). As a natural consequence of the Digital Revolution, there has been an increase in the administration

---

L.B. Russell (✉)

Centre for Health Evaluation and Outcome Sciences, Providence Health Care,  
St. Paul's Hospital, 588 – 1081 Burrard Street, Vancouver, BC V6Z 1Y6, Canada  
e-mail: [lara.russell@alumni.ubc.ca](mailto:lara.russell@alumni.ubc.ca)

A.M. Hubley

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](mailto:anita.hubley@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\_13

229

of surveys and questionnaires online, which first began to appear in the 1980s (Pew Research Center, 2011; Smyth & Pearson, 2011).

### *Advantages to Online Data Collection*

Online or web surveys offer a number of advantages over more traditional modes of data collection (e.g., interviews, paper-and-pencil surveys), including lower delivery costs, lower data collection costs per respondent, shorter turnaround times for data collection, reduced data entry time, wider range of design options, ability to conduct real-time consistency checks and other data verification, significantly more reporting of socially undesirable behaviors especially involving sensitive topics, and the ability to target some specialized or hard-to-reach populations (Fan & Yan, 2010; Fowler, 2014; Gnamb & Kaspar, 2015; Groves, 2011; Smyth & Pearson, 2011). Advances in technology have also made it easier for individuals who are not experts in web programming to design and deploy surveys online (Smyth & Pearson, 2011). Lower technical demands, combined with the affordability of online surveys, have led to what has been called the “democratization of research” (Frippiat, Marquis, & Wiles-Portier, 2010, p. 4).

### *Problems Associated with Online Data Collection*

Despite these advantages, there are problems associated with online data collection. One major concern is a lack of representativeness and the risk of bias caused by sampling issues (Bethlehem, 2010; Heiervang & Goodman, 2011). Approximately 80% of individuals in the developed world, but only 40% in the developing world, use the internet (International Telecommunications Union, 2016). Even in countries with high proportions of internet users overall, subgroups of individuals may differ in the extent to which they use the internet. In the United States, for example, individuals who are over 65, have less than a college education, lower incomes, or live in rural areas are less likely to use the internet (Anderson & Perrin, 2016). The same can be said with use of mobile devices; for example, ownership of smartphones is higher among individuals under the age of 50 and those with a college education, higher incomes, or living in non-rural environments (Pew Research Center, 2015a).

There are no databases of internet addresses equivalent to the comprehensive databases of household mailing addresses or fixed line telephone numbers that have been used in the past to ensure representative samples for postal or telephone surveys. As a result, samples for internet surveys are rarely representative of the general population (Smyth & Pearson, 2011). Probability-based internet panels do exist, but are quite rare. Instead, much online survey research relies on internet panels made up of volunteers or individuals who are compensated for their participation in surveys (e.g., Mechanical Turk). Given the differences in internet access



and sampling, it is not surprising that the samples for online surveys often differ from those obtained using other modes in terms of age, education, ethnicity, and income (e.g., Pew Research Center, 2015b). Such differences introduce a risk of bias into survey data.

There may also be differences in the quality of data collection across different types of online devices (e.g., desktop/laptop computers, tablets, mobile phones). For example, mobile phones may produce particularly problematic displays (e.g., options not visible on the screen, tables not rendered) and result in data entry errors (Callegaro, 2010; Peytchev & Hill, 2010). However, more recent research has found that missing item rates may not differ between computer and iPhone survey respondents, although the latter group completes surveys more quickly (Buskirk & Andrus, 2014).

### *Comparing Online Survey Data to Data from Paper-and-Pencil Surveys and Interviews*

The extent to which the data collected through online surveys differ from other modes is unclear. Comparisons of online and other survey modes have generally considered the following: response rates to the survey, break-off rates (i.e., not completing a survey), and psychometric properties across modes.

**Response Rates** Response rates for online surveys tend to be lower than for paper-and-pencil surveys (de Bruijne & Wijnant, 2013; Fan & Yan, 2010; Mavletova, 2013). One meta-analysis of 45 studies found response rates for online surveys to be about 11% lower than for postal or telephone surveys (Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008), although this information is based on earlier studies that may be out-of-date given the proliferation of various technologies since then. Response rates for particular modes of surveys may depend on a variety of factors such as educational level, occupational relevance, or general familiarity. For example, Boschman, van der Molen, Frings-Dresen, and Sluiter (2012) found that response rates for internet-based surveys versus paper-and-pencil surveys were not significantly different for construction supervisors (43% vs. 46%, respectively) but, for bricklayers, response rates were much lower for internet-based surveys (28%) than for paper-and-pencil surveys (44%).

**Break-Off Rates** Levels of survey break-off are higher in online surveys (as well as automated telephone interviews) compared to surveys conducted by live interview (Tourangeau, Conrad, & Couper, 2013b). Comparisons between online and postal surveys are difficult to make, as break-off rates for paper-and-pencil surveys conducted by postal mail cannot be ascertained. There are also differences in break-off rates for online surveys, depending on the devices used to take them. Surveys taken via mobile devices show much higher break-off rates compared to those taken via desktop or laptop computers or even tablets. For example, Callegaro (2010) showed break-off rates of 8.4–22.0% for desktop/laptop respondents versus 24.2–

37.4% for mobile respondents of surveys conducted in Asia, North America, and seven European countries. Wells, Bailey, and Link (2013) found break-off rates were significantly higher for smartphone respondents than for computer and tablet respondents.

**Comparability of Scores and Psychometric Properties** Response and break-off rates have generally been explored at the overall survey level. Comparisons of psychometric properties are often focused on individual scales or instruments. Studies comparing online and paper-and-pencil administration of instruments suggest that mean scores and scale reliabilities are generally similar (e.g., Alfonsso, Maathz, & Hursti, 2014; Hirai, Vernon, Clum, & Skidmore, 2011; Touvier et al., 2010; van Ballegooijen, Riper, Cuijpers, van Oppen, & Smit, 2016). However, these findings are not universal, nor are differences consistent. For example, one review found that, while mean scores were higher for internet versions of some measures, they were lower for others (van Ballegooijen et al., 2016). Lack of measurement invariance and minor differences in factor structure have also been found across modes for some measures (e.g., Buchanan, Johnson, & Goldberg, 2005; Hirai et al., 2011). One challenge with unpacking the sometimes contradictory findings is that studies comparing the psychometric properties of measures between modes often provide little information about how the measures were adapted for online administration (Alfonsso et al., 2014). This makes it difficult to determine why inter-modal differences are occurring.

### *Focusing on Online Data Collection*

Despite the challenges and unknowns associated with research using online data collection, the continual expansion of internet access and internet-enabled technology across the world and into many areas of people's lives means that online data collection is only going to increase and likely supplant traditional modes. Based on the inter-modal comparisons described above, this shift has the potential to affect the data that are obtained, though to an unknown degree. It is important for researchers to focus on understanding and evaluating elements of online measures and surveys themselves rather than attend only to comparisons to other modes. It is only when a new mode is introduced (e.g., online modes) that one starts comparing data obtained with the new mode to data obtained with the old (or standard) mode. The implication is that the new mode must measure up to the standard mode, when, in fact, there may be unrecognized or unacknowledged flaws or limitations to the standard mode. With increased use of online data collection, we need to consider the impact of *this* mode on the validity of inferences made from measures.

## Validation of Inferences Made from Online Measures and Response Processes Evidence

Validity is “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA, APA & NCME, 2014, p. 11). Can we apply validity evidence amassed on a measure that has been administered in paper-and-pencil format or in-person to the same measure that has been administered online? Has the meaning of the test scores changed as a result of using a new survey mode? Put another way, can we assume the same degree of validity for the inferences we wish to make? The answer is not clear. The inferences we make from the scores on measures are not a property of the measure per se and cannot be divorced from the respondents, the purpose of measurement, and the context in which the measurement occurs. Thus, it seems clear that it is important to obtain validity evidence to support the inferences one wants to make from measures whenever the sample, purpose, or context is different from prior research.

The *Standards for Educational and Psychological Testing*<sup>1</sup> (AERA, APA & NCME, 2014) proposes that there are five sources of validity evidence: (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing. Each of these sources can be examined in the online context. The survey mode (e.g., paper-and-pencil, online) used to present the measure affects elements such as the layout and specific features (e.g., ability to revisit items and change answers, progress indicators) that can be used (i.e., test content issues). It is possible that internal structure or relations with other variables evidence might be different because of aspects of the mode used. The focus of this book is on response processes as a source of validity evidence given the relative neglect of this source (e.g., Hubley, Zhu, Sasaki, & Gadermann, 2014; Lyons-Thomas, Liu, & Zumbo, 2014; Shear & Zumbo, 2014). Response processes refer to a series of actions, steps, or strategies that a respondent takes in selecting or providing a response to test items or stimuli. As a source of validity evidence, we are interested in whether these actions, steps, or strategies are consistent with what we would expect if we are, in fact, measuring the construct we intend.

It is worth pointing out that, in the context of online surveys, there is a large body of literature that examines how people respond to aspects – often design features – of surveys. This research examines, for example: the impact of different types of progress indicators (e.g., Conrad, Couper, Tourangeau, & Peytchev, 2010; Villar, Callegaro, & Yang, 2013; Yan, Conrad, Tourangeau, & Couper, 2011), prompts in response to skipped items (e.g., Derouvray & Couper, 2002; Leeuw, Hox, & Boevé, 2016; Oudejans & Christian, 2011), responsive grids (e.g., Couper, Tourangeau, Conrad, & Zhang, 2013; Kaczmarek, 2011), layout of response options (e.g., Toepoel, Das, & van Soest, 2009), primacy effects in responding (e.g., Galesic, Tourangeau, Couper, & Conrad, 2008), availability and methods of providing definitions or clarifications (e.g., Conrad, Schober, & Coiner, 2007; Peytchev, Conrad,

---

<sup>1</sup>Henceforth referred to as *The Standards*.

Couper, & Tourangeau, 2010), the influence of text box size on responses to open-ended questions (e.g., Emde & Fuchs, 2012; Smyth, Dillman, Christian, & McBride, 2009; Zuell, Menold, & Korber, 2015), and gamification (e.g., Harms, Wimmer, Kappel, & Grechenig, 2014; Keusch & Zhang, 2015).

Some of this research has explored the link between survey design features and resulting response behaviours and patterns. In some cases, speculations have been made about the processes underlying these patterns. For example, progress indicators that show initially rapid progress seem to reduce survey break-off rates, even if progress appears to slow later in the survey (e.g., Conrad et al., 2010; Villar et al., 2013). It is speculated that seemingly rapid initial progress may create a sense of investment in respondents (Villar et al., 2013). Adding some dynamic features to grid questions (such as making items ‘inactive’ once a response is selected) may help focus respondents’ attention and thereby reduce missing data (Couper et al., 2013; Kaczmirek, 2011). More generally, longer response times have been treated as reflective of higher attention (Emde & Fuchs, 2013), or of less stable attitudes (Heerwegh, 2003). Straightlining (i.e., selecting the same response option for all items in a grid or on a page) and speeding have been treated as indicators of inattentive, unengaged responding (Zhang & Conrad, 2013). However, these links between observed behaviours or patterns and underlying processes are speculative, and have not been explored directly. In addition, in most of this research, the constructs being measured in the survey are of little or no interest; that is, the survey is a tool to collect information about data patterns in the presence of various design features, not about a particular construct. Researchers that have focused on the psychometric properties of online measures are inherently interested in the construct being measured, but tend not to discuss the specific adaptations (e.g., design features) they have made in presenting the measures online, nor do they address how these adaptations might impact how respondents process items and select responses. Neither body of research addresses response processes in the context of online data collection as a source of validity evidence.

It seems reasonable that design features in online surveys might affect the processes involved in responding. It is also possible that the mere change of mode affects these processes. But do these changes impact the validity of inferences made from measures? And if so, how? There are many aspects of the online environment that may influence how the respondent interacts with the item and response format, potentially impacting the actions or steps that a respondent takes in providing responses, and whether those processes make sense given the construct being measured.

## **Factors Affecting the Collection of Data Online**

There are numerous factors specific to the online environment that can affect how a respondent interprets items and the steps or strategies that he/she uses to select responses or complete tasks. Some of these factors are related to how the survey is

designed, presented and delivered, while others are tied to personal factors as well as the specific technology a respondent uses to take the survey. The myriad possible combinations of these factors mean that, in some ways, no two individuals are ever taking the exact same survey.

### ***Designing and Delivering Online Surveys: Interactive and Dynamic Content***

One unique aspect of online surveys, particularly in comparison to paper-and-pencil ones, is that they can include content and features that are responsive to respondent input. For example, online surveys can be programmed to skip irrelevant items based on previous responses (using conditional branching or ‘skip logic’), provide definitions or clarifications on request (e.g., by clicking on a link), prompt respondents (e.g., if the survey is idle for a certain amount of time), and provide real-time response validation checks (e.g., warning respondents if a date has not been entered in the desired format). This interactivity can create a survey experience that, while it has much in common with self-administered paper-and-pencil surveys, also has some of the qualities of a survey administered by interview (Tourangeau, Conrad and Couper, 2013a). Indeed, the interactive nature of online surveys may be the characteristic that most distinguishes them from other survey modes.

Tourangeau et al. (2013a) discuss responsiveness in terms of dynamic features that are human-like or machine-like. Responsive features that are human-like generally have a parallel in human behaviours and may, in some cases, interact with respondents in a way that is similar to how an interviewer might. For example, a survey can be programmed to probe for additional information for open-ended questions, display prompts when a respondent chooses a non-substantive response, or display a warning when a respondent is moving through a survey too quickly. Machine-like responsive features harness the technical capabilities of computers to provide feedback in ways that are not readily available in paper-and-pencil surveys or even in interviews. For example, for questions that require cumulative responses (e.g., that should add up to 100%), a survey can provide a running tally as responses are entered. Another common feature is to program response options to change appearance or become inactive once a response has been selected, making it easier to identify unanswered items.

In addition to interactive features, online surveys use a variety of input formats using either a mouse/trackpad or touchscreen technology that are not available for paper-and-pencil surveys, such as drop-down menus, sliders, and ‘drag and drop’ item formats. Online surveys can also incorporate multi-media content, such as images, video, and sound. These increase the level to which taking a survey is a kinetic and multi-sensory experience. Another online-specific innovation is gamified surveys, which add story lines, rewards, and puzzles or mini-games to surveys and can involve both responsive and kinetic elements.

## ***Responding to Online Surveys: Technical Considerations***

Respondents primarily respond to surveys delivered via the internet using three types of devices: desktop computers, laptop computers, and mobile devices (i.e., tablets and cellular phones – primarily smartphones, though simpler surveys can be completed on cellular phones that are limited to text entry). This categorization is deceptively simple, however, as there are large variations both within and across these categories of devices that may affect the delivery and completion of online surveys.

**Screen** Screen size, aspect ratio, and resolution can differ between computer systems and mobile devices. As a result, the appearance of an online survey may differ among respondents. Larger screens allow for more visible content at any given time, which means that, depending on how the survey is designed, some respondents may see more items, response options or instructions without having to scroll down or across a screen. A larger screen also makes it easier for respondents to have multiple webpages or programs open at the same time, which might distract them from the survey itself.

**Software** In addition to the differences in screens, computers and mobile devices use different operating systems that may affect the appearance of a survey. Perhaps more importantly for online surveys, users may use different software (“web browsers”) to access the internet. Technical problems caused by compatibility issues across browsers are less common than in the past as website programming is increasingly better able to accommodate the differences among web browsers. Nevertheless, there are differences in how websites look and ‘behave’ across different web browsers that have the potential to affect a respondent’s survey experience. Computer and mobile device users can also customize their browser appearance and settings to a certain extent (e.g., changing the font size, disabling web cookies), which may affect the look and performance of a website - or online survey.

**Survey Interface** There are different ways in which respondents can input their answers and information into an online survey. Many computer users rely on a mouse and keyboard for this. Most laptops are equipped with a trackpad that can be used in place of a mouse, but this will provide a different physical experience and may be either more or less comfortable for a survey respondent to use. Tablets and smartphones often incorporate touch screen functionality, and input formats have been developed specifically to accommodate both this technology and the smaller screens of these devices (e.g., by using scrolling lists instead of drop-down menus). The differences between desktop/laptop computers and mobile devices have also led to the development of ‘mobile’ versions of many websites. These rely on different design principles than traditional websites and are optimized for small screens and touch interaction. Such differences once again change the user experience for websites and online surveys. Even if a survey is not redesigned for mobile use, accessing one from a mobile device can result in a significantly changed appearance and user experience, as the survey must resize to fit onto a small screen and

traditional forms of interaction with the survey (e.g., ‘point and click’; using the ‘Enter’ key) may be more difficult without a mouse and full size keyboard.

**Internet Connection Speed** The experience of survey respondents can also vary due to the speed of their internet connection. Websites and surveys that incorporate images, video, or large amounts of dynamic, customized or other bandwidth-intensive content will display more slowly on slower connections. Some content may not appear correctly at all. Although some regions (e.g., Europe, United States, Canada) have fairly high levels of broadband internet access, access in others is still limited (International Telecommunications Union, 2015).

**Portability** Thanks to laptops and internet-enabled portable devices such as tablets and smartphones, respondents can complete a survey in a variety of locations and environments, which may have widely different levels of distractors present.

### ***Responding to Online Surveys: Personal Factors***

The process of completing an online survey will not be influenced just by the technology involved in delivering and taking it. Personal and social context will also play a role in shaping the survey experience. Respondents’ technical skills, expectations, and attitudes towards technology in general, and digital technology in particular, all have the potential to affect respondents’ likelihood and experience of taking a survey online.

Since the 1990s, the concept of the digital divide has been part of the discussion of internet use. This divide refers to differential access to online communication and media, including internet access, connection speed, and even the quality of devices used to go online (Internet World Stats, 2016). A more recent, but perhaps more important, concept is that of ‘digital readiness’, which extends beyond access. Digital readiness refers to a broader concept of capacity – individuals may have access to the internet, but lack the knowledge, skills and confidence to make full use of online resources (Horrihan, 2014). Even in countries with high overall internet access, some individuals lack the skills to use the internet effectively. For example, approximately one-fifth to one-third of Americans have low digital skills, including poor knowledge of internet-related terms and low levels of comfort with computers, yet many of these individuals do have internet access (Horrihan, 2010, 2014).

Whether or not survey respondents are enthusiastic users of digital technology, the internet is becoming a more and more important part of many people’s lives. While not all online experiences are directly relevant to delivering or taking online surveys, the integration of the internet into daily life has the potential to contribute to expectations about how interactions with digital technology ‘should’ proceed. There are expectations around access (the internet allows access to seemingly unlimited information, content and services), availability (e.g., online banking and shopping are available 24 hours a day), and a tailored online experience (e.g., GPS-based search results, auto-filled forms). All of these may affect individuals’



expectations about their online experiences, potentially spilling over into expectations about online survey experiences. At the same time, individuals' experiences with technology will differ, so their expectations will not be the same. In fact, expectations may not even be the same for a single individual given that he/she may expect a different experience based on which device is being used. In contrast, individuals' expectations of what a paper-and-pencil survey will be like are shaped in part by the understanding that paper is a static medium. For example, when completing a paper-and-pencil survey, respondents understand that it is up to them to read and apply instructions for skipping items that are not applicable to them. In an online survey, respondents may expect the survey will automatically skip such items based on previous responses. If skip logic is poorly implemented, or not implemented at all, respondents may become frustrated – more so than when they have to work out the skip patterns for themselves for a paper-and-pencil survey. Ultimately, there are limited ways in which one can 'interact' with a paper-and-pencil survey, but enormous variety in how one can interact with online surveys.

In summary, how respondents interpret items and the response processes used to select responses can be impacted by the interactive and dynamic aspects of the online environment, technical aspects of devices used to access online surveys, and respondents' personal experiences and expectations. In the next section, we will describe how one might examine response processes as a source of validity evidence for inferences made from measures completed online.

## **Conducting Response Processes Validation Research with Online Measures**

When using response processes as a source of validity evidence, one is interested in “the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA, APA & NCME, 2014, p. 15). Put another way, we want to examine whether construct-relevant or construct-irrelevant processes are occurring. For example, for items from a self-report measure of health (e.g., ‘My health is excellent’), we would expect respondents to focus on their own health status, perhaps make comparisons to themselves at an earlier time/age or to others, and consider different aspects of health (e.g., physical, mental, pain, injury, medications). Construct-irrelevant processes might include focusing on non-health issues (e.g., “I think I’m a good person. I have good qualities.”) or not using the response format as intended (e.g., “Well, not excellent. Maybe good or very good. Strongly disagree then.”). The goal is to ensure that respondents interact with the item content and response format as intended and that the scores obtained are adequate reflections of the respondent’s experience. Ideally, if scores are not adequate reflections or response options are unable to reflect the respondent’s experience, or the response processes show a lack of relevance to the construct, one would like to

remove or avoid error variance by excluding that item or modifying the item or response options.

Cognitive Aspects of Survey Methodology (CASM; e.g., Schwarz, 2007; Tourangeau, Rips, & Rasinski, 2000) research uses a four-stage cognitive model to describe how respondents negotiate the process of responding to items, questions, or tasks. According to this model, responding to an item begins with the respondent attempting to understand the item (comprehension), then retrieving information relevant to the item from memory (retrieval), subsequently making a decision about this information within the context of the item (judgment), and finally selecting a response (reporting). Each stage of this process may involve a number of cognitive tasks.

Tasks at the comprehension stage include reading the instructions, items, and response options, understanding the meaning of the words, and determining the kind of information that is being asked. It is hoped that the respondent's understanding matches what the test developer intended but there can be problems with, for example, unfamiliar vocabulary, unclear or ambiguous wording, faulty presuppositions, and vague quantifiers. At the retrieval stage, tasks can include determining how to retrieve relevant memories, responding to item or context cues that trigger memories, and retrieving specific memories. Common errors made here include the occurrence, frequency, or timing of events in memory. As the retrieved information will rarely match the item statement perfectly, the respondent must then make one or more judgments about the completeness and accuracy of the information. If more than one relevant piece of information is retrieved, judgments must be made about whether to discard information or how to weight or synthesize the information into a single response. If information is incomplete or missing, then the respondent may have to make inferences or guess to fill in that information. Finally, the respondent must decide on, and select, a response to the item. In this stage, the respondent must format their judgment to fit the response alternatives provided in the measure. At this point, the respondent's judgment may also be edited or modified to be consistent with prior responses or as influenced by response biases (e.g., social desirability, moderate responding). As Schwarz (2007) has pointed out, CASM research attended less to the interpersonal aspects of survey interviews because meta-analyses of response effects conducted in the early 1970s indicated that task characteristics were far more influential than were interviewer and respondent characteristics (Sudman & Bradburn, 1974).

Another relevant model is Krosnick's (1991) optimizing/satisficing model. This model proposes that, while some respondents will optimize and conscientiously engage in the four CASM steps of the response process, others will provide responses that appear accurate or satisfactory but have, in fact, barely been given any conscious or unconscious thought and may skip some of the CASM stages. Some indicators of satisficing include random responding, acquiescence, primacy effects in response choice, and 'don't know' responses. Some factors that may lead to satisficing include fatigue, unclear or complex items, vague quantifiers in response options, and greater cognitive effort in responding.

A critical consideration when applying CASM and the optimizing/satisficing model in the present context is to distinguish between situations when information

and processes contribute to (a) response processes versus (b) test content, as sources of validity evidence. Recall that, for response processes as a source of validity evidence, we are interested in whether the actions, steps, or strategies that a respondent takes in selecting or providing a response to test items or stimuli are consistent with what we would expect if we are measuring the construct we intend. For test content as a source of validity evidence, we are interested in the degree to which the test elements (e.g., item content, response format, administration and scoring instructions, test layout and formatting; Haynes, Richard, & Kubany, 1995) are representative of, and relevant to, the intended construct given a particular context (i.e., purpose and sample). It is important not to confuse these two sources of validity evidence.

The fact that a survey or instrument is administered online will not necessarily affect construct-relevant and construct-irrelevant processes. In fact, we would hope that these processes are independent of the mode through which items are presented. To the best of our knowledge, there has been no research on response processes as a source of validity evidence specifically in the context of online surveys. This is not surprising, given how little response processes validation research has been conducted in general. But without such research we cannot be sure that construct-relevant response processes are unaffected by survey mode, nor can we assume that findings from research on response processes as a source of validity evidence that has been conducted in the context of other survey modes, such as paper-and-pencil surveys, apply to online surveys and instruments.

### *Using Cognitive Interviews in Response Processes Validation*

It is typically recommended (e.g., AERA, APA & NCME, 2014; Padilla & Benítez, 2014) that response processes validation research make use of cognitive interviewing (Willis, 2005), a method that uses various techniques (e.g., Think Aloud Protocols (TAP) and verbal probing; Ericsson & Simon, 1980, 1984) to study respondents' comprehension, processing, and responses to items and stimuli. Indeed, much of the limited response processes validation research has relied on cognitive interviewing (e.g., Castillo-Díaz & Padilla, 2013; Gadermann, Guhn, & Zumbo, 2011). This approach would be particularly useful in conducting research on response processes as a source of validity evidence in the context of online surveys, and some of the same technology used to deliver online surveys could be used for such direct questioning.

**Online Probing and Audio TAP** TAP and verbal probing are valuable ways to learn more about individuals' responses – for example, by asking respondents to explain why they chose particular responses. However, interview research is quite resource-intensive. In addition, it carries the risk of producing biased data for a number of reasons, including interviewer effects, the fact that interview participants tend to be disproportionately motivated, and geographically limited and/or small samples. Thus, while interviews can be useful for obtaining in-depth data from a

small number of individuals, they are not suited to collecting data from larger samples (Murphy, Edgar, & Keating, 2014). One way to conduct response processes research with a greater number of participants would be to collect some of these data in online surveys, using online probing in the form of open-ended questions. While online probing has not been applied to response processes validation research, it has been explored for other purposes, such as pre-testing survey items and investigating cross-national equivalence of survey items (e.g., Braun, Behr, Kaczmirek, & Bandilla, 2014; Murphy et al., 2014), and could be adapted for response processes validation research.

Open-ended items are generally believed to present a higher response burden than closed-ended items in surveys and are prone to higher levels of non-response (Zuelli et al., 2015). However, incorporating motivational prompts and varying the size of text boxes has been found to encourage responses to open-ended items in online surveys (Oudejans & Christian, 2011; Smyth et al., 2009). This suggests that, if properly designed and presented, online probes could be used to collect information to supplement interview data by asking some of the same questions that would be used in face-to-face interviews, and making it possible to obtain data from larger samples of participants than might be obtained through interviewing alone (see, for example, Braun et al., 2014; Murphy et al., 2014).

Nevertheless, it is important to remember that the data obtained through online probing may be different, and possibly less rich, than the data obtained through interviews. In one study comparing online probing and traditional cognitive interviews as a source of information both about problematic survey items and about respondents' reasoning behind their answers to closed-ended questions, Meitinger and Behr (2016) found that cognitive interviews had lower rates of item non-response, identified more problems with items, and produced more themes and topics in respondents' explanations for their responses. Given the relative scarcity of existing research on response processes in online data collection, detail and depth may be especially important in these early stages. Therefore, the use of online probing as a tool for research on response processes should be considered as a supplement to, not replacement for, more traditional, in-person, approaches.

Online response processes research need not be limited to written responses to probes embedded into a survey. Technology advances both on the survey creator and on the survey taker's side provide other opportunities. For example, because the majority of survey respondents cannot type as quickly as they can speak, many text responses to online probes will be edited and shortened compared to what would be obtained in an interview. As more and more computers and almost all mobile devices are equipped with microphones, one possible solution is to incorporate audio recording as a response input option for online surveys, an approach that is already used in areas like website usability testing.<sup>2</sup> Survey respondents could be presented with probes related to their response processes and would provide their responses out loud. It would even be possible to record a TAP for the entire survey, if desired.

---

<sup>2</sup>In website usability assessments, individuals record their thoughts as they navigate through and interact with a website.

Some of the potential advantages to incorporating audio responses into online response processes validation research, in addition to the ability to reach a larger number of people than through face-to-face interviews, include the reduction of interviewer effects and possibly social desirability effects, and the ability to collect data at a time and place most convenient to the participant.

**Interactive Audio Probing** Incorporating remote TAP into online surveys would not eliminate one of the most important limitations of online probing as it is currently conducted, namely the necessity of programming probes in advance and the inability to incorporate unplanned probes that arise from the content of responses. Although some amount of customization of probes is possible, we are not yet at the point where a survey can mimic an interviewer and generate highly customized probes in real time.

We are, however, currently able to implement hybrid approaches that combine online data collection with individualized contact with respondents. One example would be to have respondents complete a survey online, and then follow up with an interview via telephone or an online voice or video calling service (such as Skype, <https://www.skype.com/en/>). The interviewer would have access to the respondent's survey responses as well as passively-collected process-related paradata (e.g., keystrokes, response latencies). Certain responses or behaviours, such as particularly long or short response times or changing responses, could be flagged and explored through retrospective probing about the response processes involved.

Advantages of this approach include the opportunity for the interviewer to customize probes and seek clarification if needed, the ability to interview respondents regardless of their geographic location, respondents' flexibility to take the initial survey at a location (and possibly time) of their choosing, and the fact that the survey itself would be conducted under conditions that are closer to the typical online survey experience (i.e., not in a lab setting or with an interviewer present). However, this approach would limit sample sizes due to the necessity of one-on-one communication between interviewers and respondents. The interviews might not always be conducted soon after the survey is completed, which is when retrospective probing is most effective. Alternatively, respondents might be required to take the survey at set times when an interviewer can be available for immediate follow-up, which would impose a restriction that is not normally present for online surveys.

**Using Real Time 'Chat' Functions** In some cases, it may be desirable to probe respondents throughout a survey, rather than retrospectively. Again, existing technology could be applied to make this possible. For example, real time 'chat' functions are currently used by many companies to provide customer support and could be adapted to the research context. The respondent would take an online survey while an 'interviewer' would have access to both the survey responses and the paradata in real time. Upon observing certain responses or behaviours, the interviewer could probe for information using a standalone chat application or one embedded in the survey. Advantages of this approach include the opportunity to customize and adapt probes as needed, the lack of geographic limitations on the sample, and the interactivity of 'real time' communication. This approach shares the potential

limitation of all methods that require interviewers, in that it will limit sample sizes for practical reasons. Other potential drawbacks include its reliance on responses provided by typing, which may discourage longer responses. The perceived time pressure of the interactive component of this approach might also be uncomfortable for some respondents. However, for those who routinely use interactive text based communication, this method may feel familiar and easy.

**Using Paradata in Response Processes Validation** Although much attention is paid to the use of cognitive interviewing in collecting response processes validation evidence, *The Standards* (AERA, APA & NCME, 2014) do state that evidence about response processes can be based on eye movement, response times, successive revisions to responses, and other records of actions and steps undertaken on the way to the final response. As already noted, online surveys can permit the collection of such data, known as paradata in the online context, which include, for example, a record of mouse clicks and movements, keystrokes, how many times answers were changed for an item (and what those changes were), the number of prompts or error messages displayed, the number and timing of definitions and other help items accessed, response and inactivity times, time spent on each screen, and when a respondent has a survey open but is not interacting with it or is multitasking (Olson & Parkhurst, 2013; Sendelbah, Vehovar, Slavec, & Petrovčič, 2016). Given the particular richness of paradata from online surveys, this information can be a valuable supplement to the responses themselves, for while survey data only reflect a respondents' final answers (or non-answers) to items, paradata provide information about what happened on the way to those final answers (Heerwegh, 2011).

Paradata have been used to study respondent characteristics and behaviours, such as ambivalent attitudes, uncertainty, engagement, accessible attitudes, memory capacity, mode familiarity, guessing, and knowledge (Olson & Parkhurst, 2013). In some cases, paradata have been used to try to understand the processes behind survey responses, but without specific reference to validity and validation, and generally focusing primarily on mode-related, rather than construct-related, processes. Harnessing paradata to study response processes as a source of validity evidence would require formulating hypotheses that directly link response processes to the construct of interest and questions of validity, and then designing studies to test these hypotheses. This would bring us closer to understanding response processes in the context of online surveys. However, it is important to remember that observed response behaviours are only potential manifestations of the underlying response processes or mechanisms, and that any particular behaviour could have more than one explanation, not all of which will be related to the construct of interest and therefore to validity. For example, multiple changes to the response to an item could be due to layout of the measure or multitasking rather than a process that is directly relevant to the construct being measured. Paradata should thus be treated primarily as a way of identifying behaviours that *might* be relevant to response processes related to the construct and to validity, and therefore candidates for further exploration.

**Using Eye Tracking Data** Eye tracking technology, a specific form of paradata, has been used in survey research to investigate how individuals respond to items. For example, eye tracking data have been used to study: the amount of time that respondents spend looking at various elements of an item, with longer times spent looking at the item stem hypothesized to be indicative of problems with item comprehension (Lenzner, Kaczmirek, & Galesic, 2011); to measure the amount of time spent looking at different response options in order to study primacy effects (Galesic & Yan, 2011); and to assess respondents' use of definitions and clarifications within a survey (Galesic et al., 2008). Eye tracking has also been used alongside TAP to provide further insight into verbalizations and silences (Elling, Lentz, & De Jong, 2012; Neuert & Lenzner, 2016). Because eye tracking data can currently only be collected with specialized equipment, this research is conducted in laboratory settings. This will not always be the case. Developments in industries such as video gaming are already pushing eye-tracking technology into the consumer market, with eye tracking devices available as peripheral equipment and even built into some computers. These advances in technology for consumer devices are bringing us closer to a time when it will be possible to track eye movements for surveys conducted on personal computers and mobile devices. This will add another potentially valuable source of paradata for research on response processes as validity evidence.

None of these potential new applications of online technology will be perfect when it comes to researching response processes as validity evidence. Like all methods, they will each have advantages and disadvantages and, as with all methods, the use of these technologies will bring with them new challenges, which, while not disadvantages per se, will require changes to how research is conducted. More sophisticated applications of technology will require specialized knowledge by the research team at the design, implementation, and analysis stages. As with all studies, validity researchers would need to decide which designs and methods would be most useful - and feasible - in answering their research questions. No one study or method will be able to answer all research questions. However, making creative use of current and emerging technologies and blending these with more traditional approaches will provide us with new tools for conducting response processes validation research.

## Concluding Statements

The past few decades have seen a dramatic increase in internet use and an expansion in the types of devices that can be used to go online, opening up new opportunities for collecting information over the internet. The technology to deploy surveys and measures online is becoming more sophisticated and, at the same time, easier and less expensive to use, making online data collection an attractive option for research. However, online data collection may be affected by numerous factors related to the design of surveys, the technology used to take them, and respondents' own technology-related skills and attitudes. Because of the possible impact of survey



mode, it is important to validate inferences made from measures administered online, rather than assuming that evidence gathered with paper-and-pencil measures applies equally well to online ones.

Response processes are a source of validity evidence that has been understudied in most contexts, including online data collection. Response processes evidence could be invaluable in understanding the ways in which the online environment and characteristics of online surveys may be affecting the data collected from measures and, therefore, the inferences that can be made from these measures. At the same time, the internet provides considerable opportunities to conduct response processes validation research in new and creative ways. This chapter provides some suggestions for harnessing existing and emerging digital technologies for this purpose, in the hopes of encouraging and inspiring response processes validation research in the online context.

**Acknowledgments** We would like to thank Dr. Chris Richardson for helpful discussions on the topic of online survey research and validity.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Alfonsson, S., Maathz, P., & Hursti, P. (2014). Interformat reliability of digital psychiatric self-report questionnaires: A systematic review. *Journal of Medical Internet Research*, *16*, e268. <https://doi.org/10.2196/jmir.3395>.
- Anderson, M., & Perrin, A. (2016, September 7). 13% of Americans don't use the internet. Who are they? Retrieved from <http://www.pewresearch.org/fact-tank/2016/09/07/some-americans-dont-use-the-internet-who-are-they/>
- Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, *78*, 161–188. <https://doi.org/10.1111/j.1751-5823.2010.00112.x>.
- Boschman, J. S., van der Molen, H. F., Frings-Dresen, M. W., & Sluiter, J. K. (2012). Response rate of bricklayers and supervisors on an internet or a paper-and-pencil questionnaire. *International Journal of Industrial Ergonomics*, *42*, 178–182. <https://doi.org/10.1016/j.ergon.2011.11.007>.
- Braun, M., Behr, D., Kaczmarek, L., & Bandilla, W. (2014). Evaluating cross-national item equivalence with probing questions in web surveys. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving survey methods: Lessons from recent research* (pp. 184–200). New York: Routledge.
- Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment*, *21*, 116–128. <http://dx.doi.org/10.1027/1015-5759.21.2.115>.
- Buskirk, T. D., & Andrus, C. H. (2014). Making mobile browser surveys smarter: Results from a randomized experiment comparing online surveys completed via computer or smartphone. *Field Methods*, *26*, 322–342. <https://doi.org/10.1177/1525822X14526146>.
- Callegaro, M. (2010). Do you know which device your respondent has used to take your online survey? *Survey Practice*, *3*. Retrieved from <http://www.surveypactice.org/index.php/SurveyPractice/article/view/250>

- Castillo-Díaz, M., & Padilla, J.-L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research*, *114*, 963–975. <https://doi.org/10.1007/s11205-012-0184-8>.
- Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2010). The impact of progress indicators on task completion. *Interacting with Computers*, *22*, 417–427. <https://doi.org/10.1016/j.intcom.2010.03.001>.
- Conrad, F. G., Schober, M. F., & Coiner, T. (2007). Bringing features of human dialogue to web surveys. *Applied Cognitive Psychology*, *21*, 165–187. <https://doi.org/10.1002/acp.1335>.
- Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The design of grids in web surveys. *Social Science Computer Review*, *322*–345. <https://doi.org/10.1177/0894439312469865>.
- de Bruijne, M., & Wijnant, A. (2013). Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review*, *31*, 482–504. <https://doi.org/10.1177/0894439313483976>.
- de Leeuw, E. D., Hox, J. J., & Boevé, A. (2016). Handling do-not-know answers: Exploring new approaches in online and mixed-mode surveys. *Social Science Computer Review*, *34*(1), 116–132. <https://doi.org/10.1177/0894439315573744>.
- Derouvray, C., & Couper, M. P. (2002). Designing a strategy for reducing “No Opinion” responses in web-based surveys. *Social Science Computer Review*, *20*, 3–9. <https://doi.org/10.1177/089443930202000101>.
- Elling, S., Lentz, L., & De Jong, M. (2012). Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *IEEE Transactions on Professional Communication*, *55*, 206–220. <https://doi.org/10.1109/TPC.2012.2206190>.
- Emde, M., & Fuchs, M. (2012). Using adaptive questionnaire design in open-ended questions: A field experiment. Presented at the American Association for Public Opinion Research (AAPOR) 67th Annual Conference, San Diego, CA.
- Emde, M., & Fuchs, M. (2013). Exploring animated faces scales in web surveys: Drawbacks and prospects. *Survey Practice*, *5*. Retrieved from <http://www.surveypractice.org/index.php/SurveyPractice/article/view/60>
- Ericsson, K. A., & Simon, H. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251.
- Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, *26*, 132–139. <https://doi.org/10.1016/j.chb.2009.10.015>.
- Fowler, F. J. (2014). *Survey research methods* (5th ed.). Thousand Oaks, CA: SAGE Publications. Retrieved from <http://www.sagepub.in/books/Book231933?seriesId=Series19&publisher=%22SAGE%20US%22&sortBy=defaultPubDate%20desc&fs=1>.
- Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction with Life Scale adapted for Children: A focus on cognitive processes. *Social Indicators Research*, *100*, 37–60. <https://doi.org/10.1007/s11205-010-9603-x>.
- Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, *72*, 892–913. <https://doi.org/10.1093/poq/nfn059>.
- Galesic, M., & Yan, T. (2011). Use of eye tracking for studying survey response processes. In M. Das, P. Ester, & L. Kaczmarek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 349–370). New York: Routledge.
- Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, *47*, 1237–1259. <https://doi.org/10.3758/s13428-014-0533-4>.
- Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*, 861–871. <https://doi.org/10.1093/poq/nfr057>.
- Harms, J., Wimmer, C., Kappel, K., & Grechenig, T. (2014). Gamification of online surveys: Conceptual foundations and a design process based on the MDA framework. In *Proceedings of*

- the 8th Nordic conference on human-computer interaction: Fun, fast, foundational* (pp. 565–568). New York: ACM. <https://doi.org/10.1145/2639189.2639230>.
- Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*, 238–247. <https://doi.org/10.1037/1040-3590.7.3.238>.
- Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review, 21*, 360–373. <https://doi.org/10.1177/0894439303253985>.
- Heerwegh, D. (2011). Internet survey paradata. In M. Das, P. Ester, & L. Kaczmarek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 325–348). New York: Routledge.
- Heiervang, E., & Goodman, R. (2011). Advantages and limitations of web-based surveys: Evidence from a child mental health survey. *Social Psychiatry and Psychiatric Epidemiology, 46*, 69–76. <https://doi.org/10.1007/s00127-009-0171-9>.
- Hirai, M., Vernon, L. L., Clum, G. A., & Skidmore, S. T. (2011). Psychometric properties and administration measurement invariance of social phobia symptom measures: Paper-pencil vs. internet administrations. *Journal of Psychopathology and Behavioral Assessment, 33*, 470–479. <https://doi.org/10.1007/s10862-011-9257-2>.
- Horrigan, J. B. (2010). *Broadband adoption and use in America* (OBI Working Paper Series No. 1). Federal Communications Commission.
- Horrigan, J. B. (2014). Digital readiness: Nearly one-third of Americans lack the skills to use next-generation “Internet of things” applications. Retrieved from [http://jbhorrigan.weebly.com/uploads/3/0/8/0/30809311/digital\\_readiness.horrigan.june2014.pdf](http://jbhorrigan.weebly.com/uploads/3/0/8/0/30809311/digital_readiness.horrigan.june2014.pdf)
- Huble, A. M., Zhu, S., Sasaki, A., & Gadermann, A. M. (2014). Synthesis of validation practices in two assessment journals: Psychological assessment and the European journal of psychological assessment. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences*. London: Springer.
- International Telecommunications Union. (2015). *ICT facts & figures 2015*. Geneva, Switzerland: International Telecommunications Union. Retrieved from <https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf>.
- International Telecommunications Union. (2016). *ICT facts & figures 2016*. Geneva, Switzerland: International Telecommunications Union. Retrieved from <http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf>.
- Internet World Stats. (2016, September). The digital divide, ICT, and broadband internet. Retrieved from <http://www.internetworldstats.com/links10.htm>
- Isaacson, W. (2014). *The innovators: How a group of inventors, hackers, geniuses, and geeks created the digital revolution*. New York: Simon & Schuster.
- Kaczmarek, L. (2011). Attention and usability in internet surveys: Effects of visual feedback in grid questions. In M. Das, P. Ester, & L. Kaczmarek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 191–213). New York: Routledge.
- Keusch, F., & Zhang, C. (2015). A review of issues in gamified surveys. *Social Science Computer Review, 1–20*. <https://doi.org/10.1177/0894439315608451>.
- Krosnick, J. A. (1991). Response strategies for coping with cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213–236. <https://doi.org/10.1002/acp.2350050305View>.
- Lenzner, T., Kaczmarek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research, 23*, 361–373. <https://doi.org/10.1093/ijpor/edq053>.
- 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* (pp. 313–319). London: Springer.

- Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, *50*, 79–104.
- Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, *31*, 725–743.
- Meitinger, K., & Behr, D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods*, *28*, 363–380. <https://doi.org/10.1177/1525822X15625866>.
- Murphy, J., Edgar, J., & Keating, M. (2014). Crowdsourcing in the cognitive interviewing process. Presented at the annual meeting of the American Association for Public Opinion Research, Anaheim, CA.
- Neuert, C. E., & Lenzner, T. (2016). Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology: Theory and Practice*, *19*, 501–519. <https://doi.org/10.1080/13645579.2015.1049448>.
- Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. *Sociology Department, Faculty Publications*, (Paper 216). Retrieved from <http://digitalcommons.unl.edu/sociologyfacpub/216>
- Oudejans, M., & Christian, L. M. (2011). Using interactive features to probe and motivate responses to open-ended questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 215–244). New York: Routledge.
- Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*, 136–144. <https://doi.org/10.7334/psicothema2013.259>.
- Pew Research Center. (2011, March 18). *Internet surveys*. Retrieved April 6, 2016, from <http://www.people-press.org/methodology/collecting-survey-data/internet-surveys/>
- Pew Research Center. (2015a). *The smartphone difference*. Retrieved from <http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/>
- Pew Research Center. (2015b). *Coverage error in internet surveys*. Retrieved from [http://www.pewresearch.org/files/2015/09/2015-09-22\\_coverage-error-in-internet-surveys.pdf](http://www.pewresearch.org/files/2015/09/2015-09-22_coverage-error-in-internet-surveys.pdf)
- Peytchev, A., Conrad, F. G., Couper, M. P., & Tourangeau, R. (2010). Increasing respondents' use of definitions in web surveys. *Journal of Official Statistics*, *26*, 633–650.
- Peytchev, A., & Hill, C. A. (2010). Experiments in mobile web survey design: Similarities to other modes and unique considerations. *Social Science Computer Review*, *28*, 319–335.
- Ramasubramanian, L. (2010). The digital revolution. In *Geographic information science and public participation* (pp. 19–32). Berlin, Germany: Springer.
- Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, *21*, 277–287. <https://doi.org/10.1002/acp.1340>.
- Sendelbah, A., Vehovar, V., Slavec, A., & Petrovčič, A. (2016). Investigating respondent multi-tasking in web surveys using paradata. *Computers in Human Behavior*, *55 (Part B)*, 777–787. <https://doi.org/10.1016/j.chb.2015.10.028>.
- Shear, B., & 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). London: Springer.
- Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*, 325–337. <https://doi.org/10.1093/poq/nfp029>.
- Smyth, J. D., & Pearson, J. E. (2011). Internet survey methods: A review of strengths, weaknesses, and innovations. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 11–44). New York: Routledge.
- Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis*. Chicago: Aldine Pub. Co..

- Toepoel, V., Das, M., & van Soest, A. (2009). Design of web questionnaires: The effect of layout in rating scales. *Journal of Official Statistics*, 25, 509–528.
- Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013a). Interactive features and measurement error. In *The science of web surveys* (pp. 99–127). New York: Oxford University Press.
- Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013b). Nonresponse in web surveys. In *The science of web surveys* (pp. 37–56). New York: Oxford University Press.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
- Touvier, M., Méjean, C., Kesse-Guyot, E., Pollet, C., Malon, A., Castetbon, K., & Hercberg, S. (2010). Comparison between web-based and paper versions of a self-administered anthropometric questionnaire. *European Journal of Epidemiology*, 25, 287–296. <https://doi.org/10.1007/s10654-010-9433-9>.
- van Ballegooijen, W., Riper, H., Cuijpers, P., van Oppen, P., & Smit, J. H. (2016). Validation of online psychometric instruments for common mental health disorders: A systematic review. *BMC Psychiatry*, 16. <https://doi.org/10.1186/s12888-016-0735-7>.
- Villar, A., Callegaro, M., & Yang, Y. (2013). Where am I? A meta-analysis of experiments on the effects of progress indicators for web surveys. *Social Science Computer Review*, 31, 744–762. <https://doi.org/10.1177/0894439313497468>.
- Wells, T., Bailey, J. T., & Link, M. W. (2013). Filling the void: Gaining a better understanding of tablet-based surveys. *Survey Practice*, 6. Retrieved from <http://www.surveypractice.org/index.php/SurveyPractice/article/view/25>
- Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks: Sage Publications.
- Yan, T., Conrad, F. G., Tourangeau, R., & Couper, M. P. (2011). Should I stay or should I go: The effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. *International Journal of Public Opinion Research*, 23, 131–147. <https://doi.org/10.1093/ijpor/edq046>.
- Zhang, C., & Conrad, F. G. (2013). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8, 127–135.
- Zuell, C., Menold, N., & Korber, S. (2015). The influence of the answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review*, 33, 115–122. <https://doi.org/10.1177/0894439314528091>.

# Chapter 14

## Longitudinal Change in Response Processes: A Response Shift Perspective

Richard Sawatzky, Tolulope T. Sajobi, Ronak Brahmhatt, Eric K.H. Chan,  
Lisa M. Lix, and Bruno D. Zumbo

### Longitudinal Change in Response Processes: A Response Shift Perspective

This chapter provides a review of response shift theories, methods, and applications in patient-reported outcomes (PROs) research, with particular consideration of measurement validity and response processes. PROs refer to appraisals from patients about their self-perceived health outcomes (e.g., physical, mental, and social) that are relevant to their quality of life (QoL) (Department of Health, 2009; Fayers & Machin, 2007). Response shift is of central importance to the measurement validation of PROs because the way people interpret and respond to questions

---

R. Sawatzky (✉)

School of Nursing, Trinity Western University,  
7600 Glover Road, Langley, BC V2Y1Y1, Canada

Centre for Health Evaluation and Outcome Sciences, Providence Health Care,  
588 – 1081 Burrard Street, Vancouver, BC V6Z 1Y6, Canada

e-mail: [rick.sawatzky@twu.ca](mailto:rick.sawatzky@twu.ca)

T.T. Sajobi

Department of Community Health Sciences, University of Calgary,  
3280 Hospital Drive NW, Calgary, AB T2N 4Z6, Canada

e-mail: [tolu.sajobi@ucalgary.ca](mailto:tolu.sajobi@ucalgary.ca)

R. Brahmhatt

Ted Rogers School of Management, Ryerson University,  
350 Victoria St, Toronto, ON M5B 2K3, Canada

School of Nursing, Trinity Western University,  
7600 Glover Road, Langley, BC V2Y1Y1, Canada

e-mail: [ronak.brahmhatt@yahoo.co.in](mailto:ronak.brahmhatt@yahoo.co.in)

© Springer International Publishing AG 2017

B.D. Zumbo, A.M. Hubble (eds.), *Understanding and Investigating Response Processes in Validation Research*, Social Indicators Research Series 69,

DOI 10.1007/978-3-319-56129-5\_14

about their health and QoL may change over time. The concept of measurement validity and the processes of validation have evolved significantly over the past 20 years. A clear signal of this change can be seen in the work by Messick (1989, 1995) in his articulation of substantive validity, which focuses on evidence about the process of responding (why and how people respond) as central to measurement validation. In this context, measurement validation pertains to the process of generating evidence that is needed to justify inferences (actions and decisions) based on PRO scores. The notion of response shift, defined as “a change in the meaning of one’s self-evaluation of a target construct” (Sprangers & Schwartz, 1999, p. 1508), provides opportunity to study how response processes may change over time when a measure is repeatedly administered to the same people (Rapkin & Schwartz, 2004).

In this chapter, we focus on response shift as a phenomenon that needs to be understood in relation to response processes and measurement validation. Our purpose is to bring attention to the importance of extending beyond response shift detection and accommodation by bringing to bear the notion of response processes as validity evidence. This signals a change from *detecting* and *controlling for* response shift to an explanatory focus on *understanding* the mechanisms (mediators, moderators and other causes) by which response shift occurs. In so doing, we are recasting the conventional perspective of response shift as a confounder to consideration of the “how” and “why” of change in people’s responses, with the intent to achieve a deeper understanding of the response shift construct.

To this end, in the first section, we discuss conceptual foundations of response shift in relation to perspectives of measurement validity. As Zumbo (2007a) notes, discussions of ‘validity’ and ‘validation’ are framed and shaped by the measurement and psychometric models employed, be they classical test theory, item response theory, factor analysis, or axiomatic scaling theory. Therefore, measurement models are not neutral in the validation process (i.e., they have their own underlying values and assumptions) and their consideration is necessary for a fulsome discussion of the topic (Zumbo, 2007a, p. 54). In the second section, we review statistical methods

---

E.K.H. Chan

School of Nursing, Trinity Western University,  
7600 Glover Road, Langley, BC V2Y1Y1, Canada

Measurement, Evaluation, and Research Methodology (MERM) Program,  
The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada  
e-mail: [eric.chan.phd@gmail.com](mailto:eric.chan.phd@gmail.com)

L.M. Lix

Department of Community Health Sciences, Rady Faculty of Health Sciences,  
University of Manitoba College of Medicine, S113-750 Bannatyne Avenue,  
Winnipeg, MB R3E 0W3, Canada  
e-mail: [lisa.lix@umanitoba.ca](mailto:lisa.lix@umanitoba.ca)

B.D. Zumbo

Measurement, Evaluation, and Research Methodology (MERM) Program,  
Department of Educational and Counselling Psychology, and Special Education,  
The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada  
e-mail: [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca)



with an eye to their ability to detect, accommodate, and/or explain different types of response shift. We specifically consider how each method operationalizes response shift and to what extent each method produces information for enhancing understanding of response processes and measurement validity. The third and final section focusses on opportunities and challenges for response shift research. We specifically consider that, although response shift is inherently about response processes and measurement validation, most of the methods and applications of response shift are descriptive in nature with an eye towards ruling out potential confounding variables. There is significant opportunity for theoretical and methodological development in focusing on explanation and understanding the mechanisms (mediators, moderators and other causes) by which response shift occurs.

## Conceptual Foundations of Response Shift and Measurement Validation

Conventionally, the notion of response shift pertains to the following question that lies at the heart of the validity of inferences based on longitudinal measurements: *How can we tell whether a change in measurement scores represents a change in the attribute (or target construct) of the persons being measured or a change in the correspondence between the measure and the attribute (or target construct) at different points in time?* The concern underlying this question is that not all people will be consistent over time in *how* they interpret and respond to questions for the measurement of PROs. Longitudinal research on PROs is generally based on the assumption that the response process by which a score on a measure is produced is consistent at different points in time. To the extent that this holds true, a change in the measurement score represents a change in the corresponding attribute that is being measured. Response shift occurs when the response process, and therefore the relationship between the measure and the attribute being measured, differs over time. This, when unaccounted for, threatens the validity of inferences based on longitudinal measurements.

Initially, the concept of response shift was predominantly examined in contexts of educational training, organizational change, psychology, and management science research, where it has been broadly defined as a change in a person's internal standards of measurement. Sprangers and Schwartz (1999) extended response shift theory and research to QoL and PRO measures and defined it as a change in an individual's *internal standards, values, or conceptualizations* of the target construct (e.g., health or QoL). They theorized that response shift occurs when an individual experiences a catalyst, such as a health-related procedure or event. The catalyst influences how the individual interprets and responds to questions about health and QoL (Schwartz & Sprangers, 2009; Sprangers & Schwartz, 1999). They describe three forms of response shift (see Table 14.1) that, if ignored, could result in biased estimates of the amount of change in the construct of interest and its associations with other variables (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Oort, Visser, & Sprangers, 2009). *Recalibration* is a change in an individual's internal standards

**Table 14.1** Definitions and examples of three forms of response shift

Form of response shift	Definition	Example
Recalibration	Change in internal standards of measurement	Rectal cancer patients with a colostomy, rate QoL differently at Time 1 and Time 2 because their internal measurement scale is altered by side effects of neoadjuvant/ adjuvant treatment, the temporary nature of a stoma, and accommodation to having a stoma (Neuman, Park, Fuzesi, & Temple, 2012).
Reprioritization	Change in the values (relative importance) of domains for the measurement of a target construct	Patients with temporal lobe epilepsy over time worry more about social function compared to initial worry about seizure management (Sajobi, Fiest, & Wiebe, 2014).
Reconceptualization	Change in the definition of the target construct	Rectal cancer patients with a colostomy change their perception of what “good QoL” is; stoma-related difficulties were felt to be less important in comparison with cancer-related mortality (Neuman et al., 2012).

of the measurement scale. It pertains to situations in which an individual’s interpretation of the measurement scale changes over time. This may, for example, result in under-estimation of improvements in PROs from pre- to post-surgery (Carroll et al., 2006; King-Kallimanis, Oort, & Garst, 2010). *Reprioritization* is a change in an individual’s values toward the health construct; it manifests in a change in the relative importance of measurement indicators of the health construct. In multidimensional measures, this may include differences in the domains (e.g., physical, mental and social health domains) that are relevant to the overarching construct of interest, such as general health or QoL. If these changes in values are not taken into account, the measure of the overarching construct will be confounded by changes in the values people place on the different measurement indicators and domains by which the construct is measured. *Reconceptualization* occurs when the construct itself is redefined over time (e.g., new domains may emerge). For example, measures of QoL in contexts of palliative care often include an existential domain because this domain has shown to be of greater importance to QoL of people who have life-threatening illness (Cohen, Mount, Tomas, & Mount, 1996). When reconceptualization occurs, a new set of measurement indicators for the construct is needed to ensure that the full scope of the intended target construct is adequately represented.

Building on the above conceptual foundations of response shift and the work by Tourangeau, Rips, and Rasinski (2000) and Jobe (2003) on cognitive processes when responding to measurement items, Rapkin and Schwartz (2004) explicitly defined response shift as a change in *appraisal processes*; that is, the processes by which people conduct self-appraisals when responding to questions for the measurement of health or QoL. They view response shift as pertaining to the following four aspects of the appraisal process: (1) the “frame of reference” by which a person

responds, (2) the sampling of specific experiences that the person considers when responding, (3) the person's "subjective standards of comparison", and (4) the "combinatory algorithm" by which a person considers multiple experiences in arriving at a response to a question about their health or QoL (Rapkin & Schwartz, 2004, p. 2). The frame of reference refers to the aspects of life and previous experiences that a person considers when responding to a measurement item. This frame of reference is, in part, influenced by the content and wording of an item. The sampling of specific experiences refers to the process by which an individual considers different experiences in responding to a measurement item. The standards of comparison have to do with the types of comparisons that are made, which may include comparisons with other people, situations, or expectations. Finally, the combinatory algorithm pertains to the appraisal processes by which the various experiences are combined to arrive at a response. In other words, response shift pertains to changes in standards, values, and conceptualizations that influence the processes by which people form their subjective appraisals. It is important to note, however, that the notion of an *appraisal* implicitly reflects a process that is predominantly of a cognitive nature. This may be questioned; responses to measurement items or questions are not necessarily cognitively derived. We therefore instead use the term *response processes*. Response shift, then, refers to a change in response processes, or a change in how people interpret and respond to measurement items.

Viewing response shift from the perspective of understanding response processes directs our attention to what happens between a person's exposure to a stimulus (e.g., measurement item or question) and the moment that the person provides a response. That is, the measurement items or questions constitute the stimulus of a response. This view has several important implications for the study of response shift. First, it reminds us that changes in PRO scores should not be taken at face value but should be interpreted in relation to information about the person's frame of reference, the experiences that a person considers when responding to PRO questions, the standards of comparison, and the way that the person combines appraisals of different experiences to arrive at a response. Second, in addition to the predominant focus on the manifestation of response shift and its impact on measurement scores, it directs our attention to what happens prior to the response as the basis for understanding how the response processes themselves may have changed. Unfortunately, with few exceptions, empirical knowledge on response shift in health research has focused predominantly on what happens after the response. Consequently, knowledge about the processes of responding to PRO questions as the basis for understanding response shift is limited. Third, the view of response shift as a change in response processes necessitates that the study of response shift must, at its core, be concerned with measurement indicators (items or questions) to which people respond. Although this may seem obvious, it is noteworthy that the preponderance of PRO research on response shift has focused on the relationships between measured domains (or subscales) and their overarching constructs, instead of the examination of individual items. The distinction between "domain score shift" and "item response shift" is of particular importance because response shift at the item-level may be not be revealed when the focus is exclusively on domain

scores. For example, response shifts in different items may cancel each other out in studies that examine recalibration only at the domain level. And reprioritization of items with respect to a measured domain will not be revealed. The fourth implication is that a focus on response processes inherently situates discussions about response shift in the context of measurement validation.

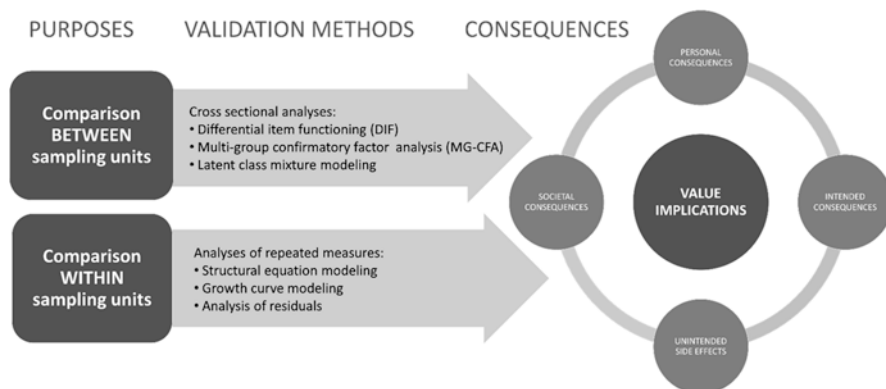
### *Response Shift in the Context of Measurement Validation*

The context of measurement validation provides an important basis for examining the implications of response shift in relation to the purposes for which we measure, the methods we use, the consequences and unintended side-effects of measurement, and, ultimately, the validity of inferences made from measurement scores. Here, we specifically draw attention to Messick's (1989) view of validity, which he defined as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). We also recognize that theories of measurement validity and validation methods have developed significantly from early research focusing on Cronbach and Meehl's (1955) ideas of a nomological network, to criterion- and content-based procedures (Anastasi, 1986), to Messick's (1989) unitary view that there is no singular source of evidence sufficient to support a validity claim. Based on these premises, "modern" perspectives on measurement validation (Cronbach, 1988; Hubley & Zumbo, 1996, 2011, 2013; Kane, 2013; Messick, 1995, 1998; Zumbo, 2007a, 2009; Zumbo & Chan, 2014) in relation to PRO measurement (Sawatzky et al., *in press*) increasingly emphasize the:

1. process by which various sources of evidence are accumulated and synthesized to support construct validity regarding the inferences and interpretations of measurement scores (including the decisions and actions based on these scores);
2. sources of evidence to establish the construct validity of measurement score inferences, including evidence based on: (1) content; (2) response processes (cognitive processes in item responses); (3) relations to other variables (e.g., convergent, discriminant, concurrent, predictive); and (4) internal measurement structure;
3. value implications of using measurement scores, including intended and unintended personal and societal consequences of measurement.

At the very core of these perspectives is the notion that we measure with a particular purpose in mind and that measurement validity therefore must be understood in relation to those purposes and the methods by which evidence pertaining to those purposes is accumulated.

The above modern perspectives of measurement validation can be further explicated in relation to three important considerations pertaining to the measurement of PROs: (1) the purposes of measuring PROs, (2) the methods by which inferences to achieve those purposes are validated, and (3) the consequences of utilizing PRO measures to achieve those purposes (see Fig. 14.1.). At a general level, the purposes



**Fig. 14.1** Purposes, Methods, and Consequences in PRO Measurement Validation

for measuring PROs have to do with wanting to make inferences based on comparisons *between* and *within* individuals or groups of people about their perceived health outcomes and QoL (Sawatzky et al., *in press*). The validity of such inferences is related to the extent that the relationships between our measures (the data we have) and their corresponding constructs (that which we wish to measure) are invariant between the sampling units of interest (individuals or groups) and over time. Thus, the measurement models and statistical methods by which the inferences based on PRO scores are substantiated are of central importance. In addition, these inferences require consideration of consequences and potential unintended side-effects (Hubley & Zumbo, 2011). That is, our inferences must be aligned with the consequences we wish to achieve while minimizing potential side-effects of measurement that we wish to avoid (e.g., biases resulting from a shift in how some people interpret and respond to PRO questions). For example, it is clear that response shift is, in part, concerned with measurement bias over time as an unintended side effect of using PRO instruments. However, there has been less attention directed towards considering response shift as an intended consequence (or outcome). In addition, understandings of personal and societal consequences of evaluating change in PRO scores have rarely been considered from a response shift point of view.

Response shift has important implications for *comparisons within sampling units* over time, particularly when groups of people are exposed to a common catalyst that induces a response shift effect, as may be the case in an experimental design. For example, Bray, Maxwell & Howard (1984) demonstrated that conventional statistical methods for the analysis of pretest and posttest data, such as paired t-tests, could have substantially lower power when response shift is present in a dataset than when response shift is absent. In fact, power losses of up to 90% were observed when the response shift effect size was moderate to large (Cohen's  $d \geq 0.50$ ). Schwartz et al. (2006) quantified the effects of recalibration in a meta-analysis of 19 intervention studies of a PRO. They found that the size of response shift effects varied across PRO domains from 0.08 (small effect) to 0.32 (moderate effect), based on Cohen's  $d$  statistic. On average, response shift resulted in down-

ward bias in estimates of change, making it more difficult to observe statistically significant treatment effects.

In addition to comparisons within sampling units over time, response shift may also affect *comparisons between sampling units* when only part of the population has been exposed to a catalyst, when there are different degrees of exposure to one or multiple catalysts, or when interactions among catalysts have occurred. In this sense, the notion of response shift has similarities with the phenomenon of differential item functioning (Zumbo, 2007b). Differential item functioning draws attention to groups of people who interpret and respond differently to measurement items or questions (i.e., measurement invariance between groups of people). Response shift may, for example, be apparent in cross-sectional health surveys that include people at different stages of illness trajectories and people who have had different degrees of exposure to health events or other potential catalysts. An example may include the comparison of people who have lived with a chronic condition to those who have not. The validity of this comparison is based on the assumption that the measurement of PROs is unbiased in relation to having lived with a chronic condition. That is, response shift could result in differential item functioning when there are different groups of people who have had different degrees of exposure to one or more catalysts for response shift.

In addition to comparisons within and between individuals or groups, response shift has implications related to considerations of both *desirable and undesirable consequences* of PRO measurement. Response shift could result in measurement bias that, if uncontrolled, will confound inferences based on, for example, a change in PRO scores. As articulated by Oort et al. (2009), “*observed* differences between respondents’ test scores may reflect something other than *true* differences in the attribute that we want to measure” (2009, p. 1126). In these situations, response shift is something that needs to be evaluated and controlled for when evaluating intervention effects. Although concern with measurement bias is paramount in response shift research about PROs, response shift can also be viewed as a form of adaptation and thus a desirable outcome. For example, response shift has been characterized as a process of homeostasis or adaptation in response to adversity (Beeken, Eiser, & Sheeran, 2008). In health care, this type of adaptation may often be viewed as a desired therapeutic effect. In other words, response shift can be viewed as a desirable change in values and experiences that influence the appraisal processes by which people interpret and respond to questions about their health and QoL. As argued by McClimans et al. (2013): “Response shifts are thus ‘evolved values’ that change in response to changed circumstances, moral development, or shifts in social standards” (2013, p. 1873). In some situations, evidence of response shift, rather than observed “true change” in PRO scores, may be the outcome of interest.

Several qualitative studies on response shift provide further insights into the relevance of response shift to measurement validation and response processes. In a study involving cognitive interviews of the Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire (EORTC QLQ-C30) (Aaronson et al., 1993) administered at multiple points in time to small-cell lung cancer patients, Westerman et al. (2008) provide the example of responses to the question “Were

you limited in pursuing your hobbies or other leisure time activities". At the start of treatment, one person selected the response option "quite a bit" with respect to his/her current hobby of gardening. Several weeks later, that person identified reading as his/her hobby and accordingly selected "a little" as the response option. As a result of a change in values, priorities or expectations, this person fundamentally changed his/her interpretation of the question about physical function. From a measurement bias point of view, this question now needs to be "recalibrated" as its position on a continuum of the degree of limitations in physical functioning has changed. Once this recalibration has occurred, the results might indicate that the person's limitations in physical function have increased, rather than decreased, in which case the person's condition will have worsened. However, from the point of view of a change in values, this person might be doing just fine (e.g., the person may have adapted to the current situation). In other words, if physical functioning is assumed to be an "objective" attribute of a person (e.g., analogous to a physiological attribute, such as height, weight, or body temperature), then response shift indicates a change in the standards by which that attribute is measured. However, if physical function is assumed to be a "subjective" attribute, then response shift is viewed as a change in social and personal values and not necessarily as a change in the "accuracy" with which the measurement item (or question) reflects the attribute being measured.

The above example reveals important insights regarding response shift in the context of PRO measurement validation. One insight pertains to the underlying motive for measuring PROs, which has to do with obtaining information about people's subjective experiences or perspectives (e.g., to complement "objective" physiological measures). However, inferences on PRO scores in health research and health services administration often reflect assumptions that are consistent with those used to evaluate outcomes that do not involve a subjective appraisal. As a result, subjective appraisals of, for example, "How do you feel?", are interpreted in the same way as objective appraisals of a person's body temperature. If the focus is on a subjective appraisal, then response shift may not necessarily constitute a form of measurement bias, but rather simply a change in the subjective appraisal. Our focus, then, is on understanding this change (e.g., a form of adaptation may have taken place). If, on the other hand, the focus is on an objective appraisal (i.e., given the same external conditions, the same score would be provided by another person or at a different point of time), then response shift becomes a source of measurement bias that needs to be controlled for. The example further illustrates the importance of understanding the values and perspectives by which people respond to questions of relevance to their QoL. These values and perspectives may change over time. Overall, we are reminded that PRO scores should not be taken at face value; there is a need for theoretical understandings that explain how and why a change in PRO scores has occurred. This understanding is critical to the validity of inferences based on PRO scores. In addition to qualitative research and philosophical analysis, empirical research involving the use of statistical methods is needed to detect, accommodate, and explain response shift.



## Statistical Methods for Examining Response Shift

The notion of response shift in PRO measurement has given rise to a variety of statistical methods to detect, control for, and explain the occurrence of response shift. To some extent, the different perspectives of response shift in the context of measurement validity are framed and shaped by the different statistical methods that have been employed. Quantitative methods for examining response shift can be classified as consisting of (a) *design-based approaches* that involve inclusion of strategies in the design of the study that have the specific purpose of detecting or controlling for response shift and (b) *model-based approaches* that involve the use of statistical methods for testing response shift hypotheses without changing the design of a study (see Table 14.2).

**Table 14.2** Quantitative methods for examining response shift

<b>Design-based approaches</b>	<b>Examples</b>
<i>Retrospective pre-test design</i> <b>Uses:</b> To detect and accommodate for recalibration response shift.	Then test (Schwartz & Sprangers, 2010)
<i>Individualized measures</i> <b>Uses:</b> To detect and accommodate for reprioritization response shift.	Patient Generated Index (Ruta, et al., 1994) Schedule for Evaluation of Individual Quality of Life (SEIQOL) (Hickey et al., 1996)
<b>Model-based approaches</b>	<b>Examples</b>
<i>Latent measurement modeling methods</i> <b>Uses:</b> Primary used to detect any form of response shift; however, there is potential to accommodate for response shift and the methods can be extended to include exogenous explanatory variables.	Domain-level latent variable models using structural equation modeling (Oort, 2005) Item-level latent variable models including structural equation modeling (Verdam, Oort, & Sprangers, 2016) and item response theory (IRT)/Rasch (Guilleux et al., 2015)
<i>Longitudinal mixed-effects regression</i> <b>Uses:</b> Primarily used to detect response shift and identify individuals who experienced response shift; however, the form of response shift may not be readily deduced.	Mixed effects regression models with interactions (Bernhard, Lowy, Maibach, & Hurny, 2001) Group-based trajectory modeling of mixed-effects regression residuals (Mayo et al., 2008)
<i>Relative importance analysis</i> <b>Uses:</b> Primary use is to detect reprioritization response shift by testing for changes in relative importance of QoL domains. The methods could be used to test explanatory hypotheses.	Importance measures based on logistic regression and discriminant analysis (Lix et al., 2013) Variable importance based on random forest regression (Boucekine et al., 2013)
<i>Classification/Data mining techniques</i> <b>Uses:</b> To detect any form of response shift.	Recursive partitioning (Li & Rapkin, 2009)

## ***Design-Based Approaches***

Design-based approaches require that methods for examining response shift are incorporated into the research design a priori. These include the use of the then-test and individualized PRO measures.

**Then-Test** The most commonly adopted method for identifying response shift is the then-test, which is a retrospective pre-test design in which PRO measurement is conducted both before the intervention/catalyst (pre-test) and after the intervention/catalyst (post-test) (Schwartz & Sprangers, 2010). During the post-test, a retrospective “then” measure is administered by asking participants to report their current appraisal of the measure at pre-test. The argument is that the difference between the “then” measure and the pre-test measure is representative of response shift (predominantly recalibration response shift). Accordingly, the difference between “then” measure and the post-test measure is viewed as a measure of the treatment effect that is adjusted for response shift. It is noteworthy, however, that the then-test was originally designed to be used in the context of experimental designs where response shift is indicated when the differences between the “then” measure and the pre-test measure are also different within the experimental groups (i.e., the intervention is the catalyst for response shift). Interpretation is more challenging when the then-test is applied to different research designs that do not involve both within- and between-group comparisons on the intervention (i.e., the potential catalyst for response shift).

Despite the frequent use of this approach, there are significant methodological and conceptual issues that may confound the results. The most obvious issue is that then-test results may be biased due to inaccurate recall and social desirability bias (Nolte, Elsworth, Sinclair, & Osborne, 2009). In addition, the then-test is based on the assumption that a comparison of then-test and post-test scores ought to be equivalent to the comparison of pre-test and post-test scores minus response shift (i.e., they have a common internal standard of measurement). In other words, the appraisal of the memory of the experience is assumed to be conceptually the same as the appraisal of the experience at the present time. As shown by Schwartz and Rapkin (2012), there are good reasons to believe that this assumption may not be warranted. Although the then-test may reflect response shift, there are often other potential explanations for discrepant results.

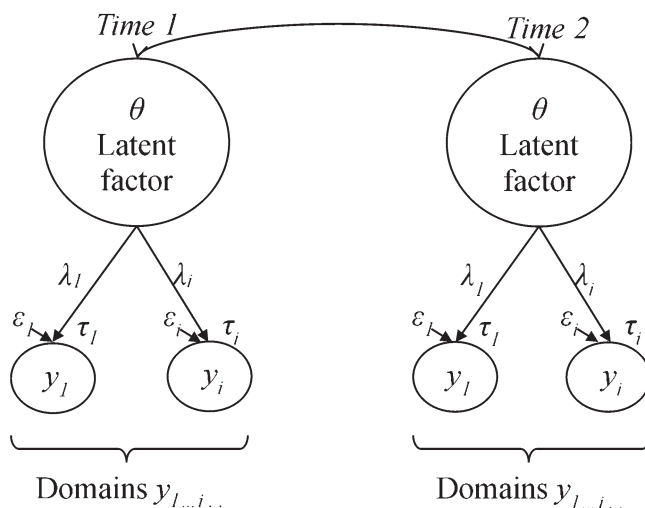
**Use of Individualized PRO Instruments** Most commonly used PROs have pre-defined sets of items. One of the assumptions of such instruments is that the same set of items and the importance of them are constant over time. Although the items included in most of these instruments are based on rigorous psychometric evaluation and results from patient interviews, this approach to PRO measurement neglects the potential changes or dynamics of QoL that people experience throughout disease trajectories (Ahmed, Mayo, Wood-Dauphinee, Hanley, & Cohen, 2005). For instance, patients with arthritis may, at the beginning, be concerned about the pain

associated with the disease but may, later on, be more concerned about their ability to participate in leisure activities.

Individualized PRO instruments have the benefit of allowing the use of a set of predetermined items to evaluate if people's conceptualization of QoL or the importance of QoL domains change over time. The Patient Generated Index (PGI) (Ruta et al., 1994) and the Schedule for Evaluation of Individual Quality of Life (SEIQoL) (Hickey et al., 1996) are two widely used individualized PROs (Martin, Camfield, Rodham, Kliempt, & Ruta, 2007). Patients completing these instruments are asked to select, rate, and weight the relative importance of areas of their life that are of greatest importance or relevance to their QoL. The PGI, for instance, consists of three stages. In the first stage, patients are asked to nominate five important areas of their life most affected by their health condition. In the second stage, each of the five nominated areas is rated on a 10-point scale, with 0 representing "the worst you can imagine" and 10 representing "exactly as you would like to be." Patients are also asked in the second stage to rate a sixth area meant to cover "all other areas of life not already mentioned." In the final stage, patients indicate the relative importance (i.e., weight) of each of the nominated areas for their overall QoL. A single index score is subsequently generated (Martin et al., 2007). When used at different points in time, individualized PROs, including the SEIQoL and the PGI, allow for the investigation of reconceptualization and reprioritization response shift by evaluating changes in the occurrence and relative importance of each identified area (Ahmed et al., 2005).

### ***Model-Based Approaches***

Several model-based approaches for the examination of response shift have been developed during the past two decades. These approaches are particularly useful for secondary analyses of response shift in longitudinal data where a priori design-based approaches have not been utilized (Schwartz et al., 2013). Model-based approaches can be broadly classified according to the following types of methods: (a) latent variable measurement modeling (e.g., structural equation, item-response theory, and latent class mixture models) (Barclay-Goddard, Lix, Tate, Weinberg, & Mayo, 2009; Oort, 2005; Oort, Visser, & Sprangers, 2005; Schmitt, 1982; Schmitt, Pulakos, & Lieblein, 1984), (b) the use of mixed effects regression models that include random-effects regression with interactions (Lowy & Bernhard, 2004) and group-based trajectory modeling of mixed-effects regression residuals (Mayo et al., 2008), (c) relative importance analysis (Lix et al., 2013; Sajobi et al., 2012) and random forest analysis (Boucekine et al., 2013), and (d) classification or data mining techniques, such as recursive partitioning (Li & Rapkin, 2009). With respect to the latent variable measurement modeling methods, it is important to note that analyses of response shift in PRO research have, thus far, predominantly focused on the relationships between summary scores (or PRO domain scores) and a total score (e.g., overall health or QoL). Although new methods have recently been



**Fig. 14.2** Factor analysis specification of a response shift model at the domain level

*Notes.*  $\lambda$  = factor loadings for domains  $y_i$ ,  $i = 1, \dots, I$ .  $\tau$  = intercepts for domains  $y_i$ ,  $i = 1, \dots, I$ . The same domains are correlated across both time points (not shown). Hypothesis of reprioritization response shift:  $\lambda_i$  (time 1)  $\neq$   $\lambda_i$  (time 2). Hypothesis of recalibration response shift:  $\tau_i$  (time 1)  $\neq$   $\tau_i$  (time 2)

proposed for examining response shift at the item level (Ahmed, Sawatzky, Levesque, Ehrmann-Feldman, & Schwartz, 2014; Anota et al., 2014; Guilleux et al., 2015; Verdam et al., 2016), they have not yet been widely-used. As a result, response shift with respect to the response processes at the item level has not been extensively examined in PRO research.

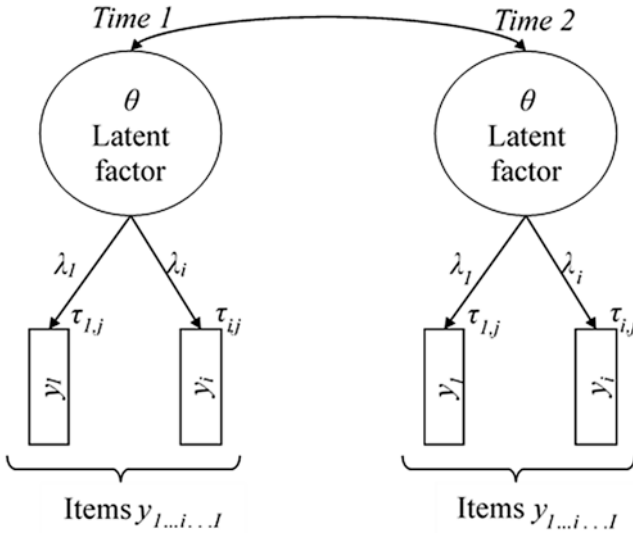
**Domain-Level Latent Variable Models** The latent variable measurement modeling approaches involve assessment of response shift with respect to measurement equivalence of latent variables over time. Specifically, these models are used to test the extent to which measurement model parameters that define the relationships between PRO domain scores (or summary scores) and the latent variable are consistent (or invariant) over time (see Fig. 14.2).

This class of methods includes structural equation modeling (SEM), item-response theory models (IRT), and latent variable mixture models. SEM is a multivariate technique that combines confirmatory factor analysis and regression analysis to investigate measurement equivalence of latent variables over time in PRO studies. Schmitt (Schmitt, 1982; Schmitt et al., 1984) proposed the use of covariance structure analysis with SEM to identify both reconceptualization and recalibration response shift. However, this approach does not identify reprioritization or non-uniform recalibration response shift. Oort (2005) later expanded the SEM approach to identify all of the types of response shift. While Schmitt's and Oort's approaches for identifying response shift are different, they are both based on tests of model parameters to detect response shift. In Oort's approach, response shift is tested

using a four-step procedure that relies on the comparison of fit statistics for models in which parameter estimates are constrained to be equivalent across measurement occasions or allowed to remain unconstrained. The interpretation of the change in patterns or magnitude of the model parameters is associated with different types of response shift. For example, a change in the pattern of factor loadings between two measurement occasions indicates the presence of reconceptualization response shift, while changes in the magnitude of the factor loadings over time indicate the presence of reprioritization response shift. Uniform recalibration is indicated if differences exist between item intercepts while constraining common factor loadings to be equivalent. Non-uniform recalibration is said to be present when there are differences in item residual variances over time. Oort's approach continues to be the most frequently used SEM method for detecting response shift.

**Item-Level Latent Variable Models** Until recently, SEM approaches for response shift detection were conducted exclusively on subscales of multidimensional instruments, using linear models. That is, item-level distributions were not included in the response shift models. This is a significant limitation considering that the notion of response shift has to do with how people appraise and respond to items. Although analyses at the item-level are more desirable, a significant challenge exists in that the item distributions are typically of an ordinal discrete nature. Methods for item-level response shift analysis have been proposed to address this limitation. Ahmed, Sawatzky and Schwartz (2014) describe a model that involves representing the relationship between discrete indicator variables and the latent factors in the form of a probit link function. The authors first examined response shift with respect to these relationships and subsequently fit a second-order latent factor model to examine response shift between these previously specified latent factors (each representing a dimension of a multidimensional measure) and the overarching construct. The model is fit using least-squares estimation. Verdam, Oort and Spranger (2016) describe a similar two-step approach. The first step is to test the relationship between each observed ordinal discrete variable and its corresponding item-specific continuous latent variable. In doing so, each item is represented by a continuous latent variable by fitting the ordinal response options under a standard normal distribution (i.e., equivalent to the use of polychoric correlations). The second step involves evaluating response shift with respect to the item-specific continuous latent variables and their corresponding latent factors.

In addition to the SEM approaches described above, item response theory (IRT) has been proposed for examining response shift at the item level. IRT involves specifying the relationships between discrete items and the latent factors in terms of a logistic link function. In a 2-parameter IRT model, such as Samejima's (1997) graded response model, this relationship is defined in terms of a proportional-odds logistic regression (see Fig. 14.3). As in the SEM approach, tests of specific measurement model-parameter constraints can be conducted to detect corresponding forms of response shift, as is described in Fig. 14.3. Although different IRT models for examining response shift have been demonstrated in educational testing and



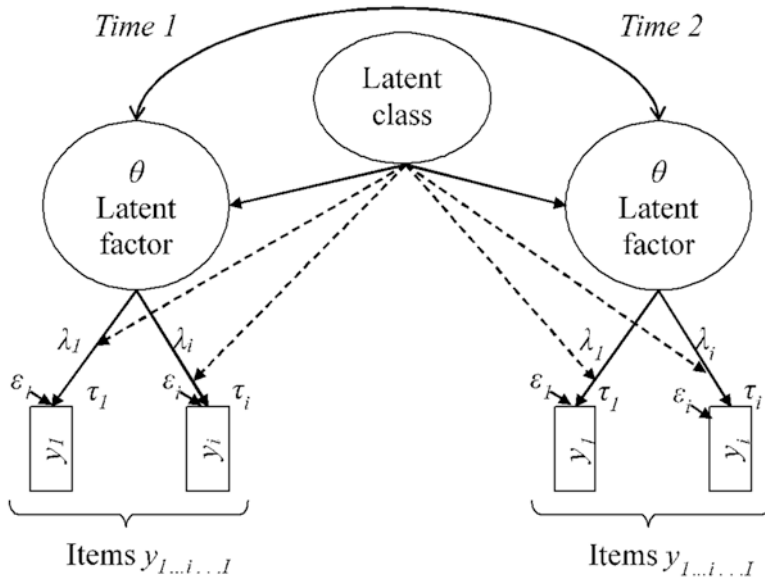
**Fig. 14.3** Item response theory specification of a response shift model  
*Notes.*  $\lambda$  = factor loadings for items  $y_i$ ,  $i = 1, \dots, I$ .  $\tau$  = thresholds for  $j - 1$  response categories per item. The same items are correlated over time. The cumulative probability  $P_{ij}$  of an item  $i$  response at or above category  $j$ :

$$P_{ij}(Y \geq j|\theta) = \frac{\exp(-\tau_{ij} + \lambda_i\theta)}{1 + \exp(-\tau_{ij} + \lambda_i\theta)}$$

Hypothesis of repriorization response shift: the factor loadings (discrimination parameters) are not invariant over time:  $\lambda_i(\text{time } 1) \neq \lambda_i(\text{time } 2)$ . Hypothesis of recalibration response shift: the thresholds (difficulty parameters) are not invariant over time:  $\tau_{ij}(\text{time } 1) \neq \tau_{ij}(\text{time } 2)$

developmental psychology (Kim & Camilli, 2014; Meade, Lautenschlager, & Hecht, 2005; Millsap, 2010), there few examples pertaining to the measurement of PROs. A notable recent example is a study by Guilleux et al. (2015), which illustrates the use of the generalized partial credit model to examine recalibration and reprioritization response shift.

A particular advantage of the latent variable models for response shift detection is that they can be expanded to examine potential sources of response shift or group differences in response shift. For example, King-Kallimanis and colleagues (King-Kallimanis, Oort, Nolte, Schwartz, & Sprangers, 2011; King-Kallimanis, Oort, Visser, & Sprangers, 2009) demonstrate a model that involves the inclusion of explanatory (exogenous) variables to evaluate whether response shift is reflected in the relationship between the outcome of interest and explanatory variables. They refer to this as “explanation bias response shift”. And Lix et al. (2015) demonstrate a multigroup modeling approach to detecting differences in response shift across different groups. This approach significantly contributes to the field of response shift research by offering a method for examining factors that may explain the occurrence of different forms of response shift. This is done by allowing the response



**Fig. 14.4** Latent class model for response shift detection

*Notes.* --- → Allow the measurement model parameters to vary across two or more latent classes

shift effects to vary across two or more groups and by testing for both between-group and within-group measurement invariance. Sawatzky et al. (2012) have further demonstrated how this approach can be expanded to situations where relevant grouping characteristics are not known a priori; the response shift model is nested within two or more latent classes (see Fig. 14.4). Although this approach is experimental, it provides promising opportunity for further development, including the ability to identify different classes of people who experience different degrees or forms of response shift.

To summarize, latent variable measurement modeling approaches are advantageous in that they can be used to test for specific hypotheses about different forms of response shift. The process involves sequentially testing different model constraints. They can be used to evaluate response shift at both the item-level and the domain-level. However, although some guidelines exist, it is often unclear exactly which constraints are more tenable than others (i.e., nearly identical, and sometimes equivalent, results could be obtained when comparing competing models that involve different constraints). There is therefore some ambiguity regarding accurate identification of response shift. In addition, due to the fact that often many model parameters are estimated, latent variable models typically require large sample sizes to achieve optimal statistical results (e.g., stable and estimable model parameters). Furthermore, the impact of response shift effects may be masked. The method may also lack statistical power to detect response shift in studies where only a small proportion of the sample experiences response shift.



**Longitudinal Mixed Effect Regression Analysis** This class of methods relies on the modeling of longitudinal change in a PRO measure and possible explanatory variables while accounting for within-subject variation and between subject variations through random effects models. Lowy and Bernhard (2004) proposed a *random-effects regression model*, in which the longitudinal change in global PRO outcome is modeled as a function of the component domains, and domain-time interactions. In this methodology, reconceptualization response shift is described as evidence of change in the relationship between the target construct and component domains moderated by time. Statistical significance of the domain-time interaction effect, which is interpreted as the changing effect of the domain on the global construct over time, indicates the presence of reconceptualization response shift. The presence of reconceptualization response shift among all of the domains can be tested as a global null hypothesis that the regression parameters for all of the domains remained constant over time using a likelihood ratio test. Following the rejection of this global null hypothesis, parameter estimates and standard errors of the interaction terms for each domain are used to determine domains for which the interactions terms are significant.

Despite the relative ease of implementing this method in most of the existing statistical software packages, this method has some limitations. First, it can only be applied to test for response shift in PRO measures with component domains and a global/summary score. It may not be applicable in measures with no global scores. Second, given that PRO measures are often characterized by strongly-correlated domains, the inclusions of the correlated domains as explanatory variables may make the random effects regression model parameters and standard errors susceptible to the effects multi-collinearity.

Mayo et al. (2008) also proposed a latent-trajectory model based on *random-effects models*. This method conceptualizes response shift as changes in unexplained variation in a PRO measure over time, after adjusting for potential determinants and covariates (i.e., residuals). Specifically, a random-effects regression model is used to model the longitudinal change in a PRO measure over time using time varying covariates. Then a group-based latent trajectory model is applied to correlated residuals derived from the random-effects regression model to identify latent subgroups of subject-specific trajectories. Subjects with changing residual trajectories are considered to experience response shift. This method is particularly advantageous because it can be used to detect response shift at the individual level. However, this method has a number of potential limitations. First, the implementation of a group-based latent trajectory model on the regression residuals often requires data on each participant on at least four measurement occasions. Second, this methodology may be sensitive to model misspecification, especially when data are not collected on important explanatory variables.

**Relative Importance Analysis** For this class of methods, response shift is operationalized as changes in the relative importance of the PRO domains over time. Here the relative importance weight attributed to a domain is not elicited based on patient direct report, but based on the marginal statistical evaluation of the relative impor-

tance of the domains with respect to a target construct (e.g., global PRO construct or response shift catalyst indicator). Specifically, measures of variable importance based on a regression modeling framework (Sajobi et al., 2012; Thomas, Hughes, & Zumbo, 1998), discriminant analysis (Thomas, 1992; Thomas & Zumbo, 1996), and recursive partitioning models (Breiman, 2001) are used to derive the rank order of the domains of a PRO measure with respect to a target construct (e.g., disease severity, disease diagnosis, or type of treatment received).

Lix, Sajobi, Sawatzky et al. (2013) developed tests based on change in the relative importance of a PRO domain, which is considered to be the ability of a domain to distinguish between groups, over time, to test hypotheses about reprioritization response shift. The importance of each domain is estimated based on magnitude and ranks of relative importance weights derived from discriminant analysis (Huberty & Wisenbaker, 1992; Thomas, 1992; Thomas & Zumbo, 1996) or logistic regression models (Thomas et al., 1998; Thomas, Zhu, Zumbo, & Dutta, 2008). Specifically, relative importance weights correspond to functions of logistic regression coefficients and discriminant function coefficients. These include standardized regression/discriminant function coefficients, Pratt index for logistic regression, and discriminant ratio coefficients, which have been extensively reported in the literature of discriminant analysis and logistic regression. Although these measures quantify the relative importance of an explanatory variable (e.g., domain), there are subtle differences in their conceptualization of domain importance. While measures of relative importance based on discriminant analysis provide information on a variable's importance in its contribution to multivariate group separation, measures of relative importance based on logistic regression analysis provide information about a variable's importance in terms of its independent contribution to group discrimination (i.e., explained variation in group membership). Reprioritization response shift in a domain is tested by estimating the distribution of changes in relative importance of a domain between two measurement occasions via a bootstrap distribution. Reprioritization response shift is considered present in a domain if the observed change in relative importance weights and/or ranks is greater than  $100(1 - \alpha)$  percentile of the bootstrap distribution of the differences in relative importance weights/ranks.

One limitation of the approach by Lix et al. (2013) is that measures of relative importance based on discriminant analysis and logistic regression are known to be sensitive to moderate to strong domain correlations, which may lead to incorrect rank ordering of the domains. Hence, the tests may be less powerful in detecting reprioritization response shift when the data exhibit strong domain correlations. On the other hand, given that tests of hypothesis about reprioritization of multiple domains may inflate the overall family-wise Type I error rate, Lix et al. (2013) suggested the use of a Bonferroni correction of the overall type I error. Although this methodology was originally developed for assessing reprioritization response shift of PRO domains with respect to group differences, this methodology can also be used to assess the reprioritization response shift in PRO domains with respect to a continuous target construct (e.g., global or summary score).

Boucekine, Loundou, Baumstarck et al. (2013) investigated the use of variable importance measures derived from random forest regression, which is an ensemble method derived by repeatedly conducting recursive partitioning of the data in homogeneous groups (classification and regression tree) via bootstrap sampling from the original data (Breiman, 2001). This approach was first used to test for reprioritization response among PRO domains in a secondary longitudinal data set of multiple sclerosis patients. Unlike the relative importance approaches proposed by Lix et al. (2013), which can only test for reprioritization between two measurement occasions, the random forest regression approach can be used to test reprioritization response shift over more than two measurement occasions. In random forest regression, the importance of a domain is quantified using the average variable importance (AVI), which is estimated as a mean relative decrease in the trees' performance when the observations of this variable are randomly permuted. The presence of reprioritization response shift among the domains can be examined graphically by plotting the domain-specific AVIs against time. Reprioritization response shift is considered to be present among the domains if the AVI curves for two or more domains intersect.

Random forest regression may lead to erroneous rank ordering of the PRO domains at each measurement occasion, especially when the classification and regression tree models developed for the bootstrap samples are least parsimonious (i.e., model over fitting). This can lead to wrong conclusions about the presence of reprioritization response shift among the domains over time. On the other hand, given that the AVI for the PRO domains are computed independently at each occasion, the assessment of changes in the domain-specific AVI over time ignores the domain-specific correlation over time.

**Classification/Data Mining Techniques** Li and Rapkin (2009) proposed the classification and regression tree (CART) method, a data mining technique, to detect response shift. This method uses non-parametric techniques to iteratively partition the data into increasingly homogeneous subgroups and then implements a prediction or regression model within each partition in a way that maximizes the explained variance within each subgroup. Consequently, the partitioning of the data can be represented as a decision tree. Using CART, change in a PRO outcome (e.g., a global summary score) between two occasions can be modeled as a function of the baseline PRO outcome score and other relevant predictors (e.g., relevant domains) of longitudinal change in the outcome across different subgroups (e.g., disease severity or treatment groups). Recalibration response shift is considered present when relationships between predictors and outcome scores utilize different group-specific cut-off points to identify homogeneous groupings over time. Reprioritization response shift is inferred based on changes in the order of the PRO domains in CART tree pathways over time, while reconceptualization response shift is inferred based on changes in the content and/or number of domains by group in the tree over time. The CART method is particularly advantageous because it makes few assumptions about the data but its use is highly exploratory in nature. Additionally, the use of CART for response shift detection is prone to misclassification error, which

limits generalizability of the findings (unless validation is done in additional independent samples).

## **Opportunities and Challenges**

Our review of conceptual foundations and statistical methods brings to light the potential of research on response shift to further our understanding of response processes, particularly in relation to the longitudinal measurement of latent constructs. The most important, and obvious, opportunity pertains to the notion that response processes are not necessarily static and may change over time when particular events or life circumstances cause people to think differently about the construct that is being measured. In those situations, the possibility of response shift needs to be taken into account when making inferences based on comparisons of PRO scores over time or between different groups of people who have been differentially exposed to response shift catalysts. There have been significant advances in the development of complex statistical methods for examining response shift. To date, these methods have been predominantly applied for purposes of response shift detection. A few statistical methods, notably the latent variable methods, can be used to obtain PRO scores that accommodate, or adjust for, response shift. Statistical methods can also be used to identify catalysts of response shift, identify people who are prone to experiencing different forms of response shift, and evaluate variation in response processes over prolonged periods of time. Knowledge about potential catalysts for response shift and the characteristics of people who experience response shift is useful for the development of explanatory theories about the occurrence of different response processes. In addition, evaluations of response shift over longer periods of time could help to determine whether response shift represents a permanent change in response processes or whether the change is temporary. However, despite the potential of research on response shift in PROs to investigate explanatory propositions, most of the studies to date have focused on the detection of response shift and, to some extent, accommodation of response shift when evaluating changes in PROs over time.

Notwithstanding the opportunities for response shift research to contribute to knowledge development about response processes, there are important limitations and challenges that point toward the need for ongoing methodological development. A particular concern is that most current statistical methods for response shift detection are not explicated in relation to theories of response processes and measurement validation. Without this foundation, the choice of statistical method may be primarily guided by technical considerations, such as the design of a study, sample size, variable distributions, and so on. However, the various statistical methods operationalize response shift in different ways and highlight different types of response shift. It is, therefore, important that decisions about which response shift

method to utilize be informed by substantive considerations. From a measurement validation point of view, these considerations should be grounded in the particular purposes of using PRO information and the types of actions and decisions that will be made based on the PRO information. For example, if the primary purpose is to adjust for response shift in a large clinical trial, then a method that produces adjusted response shift scores should be utilized (e.g., SEM or IRT). If the purpose is to understand how people are interpreting and responding to items, then methods that allow for item-level analyses should be utilized. On the other hand, if one wishes to understand how the values regarding QoL domains may change over time, then relative importance approaches to response shift detection would be most important.

In addition, there is a need for ongoing methodological advances to address issues pertaining to distributional assumptions, missing data, and the analysis of response shift at the item level. Statistical methods available for examining response shift are mostly parametric methods, which are based on certain assumptions and are not equally sensitive to data characteristics (e.g., collinearity, missing data). These assumptions are often violated in PRO data, which are generally characterized by skewed or heavy-tailed distributions, incomplete data, and strong correlations among domains and items (Beaumont, Lix, Yost, & Hahn, 2006). Articles on response shift often do not report on the assumptions underlying a response shift method and how they are tested. To date, there has been limited research on the consequences of violations of statistical assumptions. For example, in a special issue on the impact of missing data on response shift detection published in the journal *Quality of Life Research*, a number of studies showed that SEM, item response theory and relative importance analysis methods may be less powerful in detecting response shift effects in incomplete longitudinal data (Guilleux et al., 2015; Sajobi et al., 2015; Schwartz et al., 2015; Verdam, Oort, van der Linden, & Sprangers, 2015). Further methodological research is needed to assess the implications of departures from statistical methods' distributional assumptions on the detection of response shift. Simulation studies may be particularly valuable in this respect.

Finally, statistical methods for examining response shift have often been applied to secondary analyses of existing data. Theoretical understandings of responses shift could be better advanced by including the investigation of response shift a priori in the design and implementation of PRO research. These may include mixed-methods approaches that combine statistical analysis with qualitative information. Qualitative research could help to review the differences and changes in the frame of reference, values and conceptual perspectives by which people respond to questions about their health and QoL. Once further developed, measures of these processes can be included the research design. The QOL Appraisal Profile instrument by Rapkin & Schwartz (2004); Rapkin et al. (2016), which was designed to measure specific aspects of the appraisal process, is particularly promising in this regard. More research of this nature is needed to advance theoretical understandings of response shift that are grounded in theories of response processes and measurement validation.

## Conclusion

Our expository review of conceptual and methodological underpinnings of response shift has revealed several areas for further theoretical and methodological development. Response shift is, at its core, a phenomenon that pertains to underlying response processes. Although theoretical premises emphasize the importance of understanding the processes that lead to the response shift, empirical research on response shift has predominantly been concerned with the responses themselves (i.e., the PRO scores). Understanding of longitudinal change in response processes is integral to the validity of inferences pertaining to changes in PROs over time. There is significant potential for further development of theoretical understandings of response shift by researching what goes on between the stimulus (the presentation of a measurement item) and the response.

**Acknowledgements** Writing of this chapter was supported by a CIHR Operating grant (#342467) held by RS, TTS, LML, and BZ; a Research Manitoba operating grant held by LM; an O'Brien Institute for Public Health Catalyst award held by TTS; Canada Research Chairs program funding for a Canada Research Chair in Patient-Reported Outcomes held by RS; a Manitoba Research Chair held by LL.

## References

- Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, *85*, 365–376.
- Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A., & Cohen, S. R. (2005). The structural equation modeling technique did not show a response shift, contrary to the results of the then test and the individualized approaches. *Journal of Clinical Epidemiology*, *58*, 1125–1133. doi:10.1016/j.jclinepi.2005.03.003.
- Ahmed, S., Sawatzky, R., Levesque, J. F., Ehrmann-Feldman, D., & Schwartz, C. E. (2014). Minimal evidence of response shift in the absence of a catalyst. *Quality of Life Research*, *23*, 2421–2430. doi:10.1007/s11136-014-0699-3.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*, 1–15.
- Anota, A., Bascoul-Mollevis, C., Conroy, T., Guillemin, F., Velten, M., Jolly, D., et al. (2014). Item response theory and factor analysis as a mean to characterize occurrence of response shift in a longitudinal quality of life study in breast cancer patients. *Health and Quality of Life Outcomes*, *12*, 32. doi:10.1186/1477-7525-12-32.
- Barclay-Goddard, R., Lix, L. M., Tate, R., Weinberg, L., & Mayo, N. E. (2009). Response shift was identified over multiple occasions with a structural equation modeling framework. *Journal of Clinical Epidemiology*, *62*, 1181–1188. doi:10.1016/j.jclinepi.2009.03.014.
- Beaumont, J. L., Lix, L. M., Yost, K. J., & Hahn, E. A. (2006). Application of robust statistical methods for sensitivity analysis of health-related quality of life outcomes. *Quality of Life Research*, *15*, 349–356. doi:10.1007/s11136-005-2293-1.
- Beeken, R., Eiser, C., & Sheeran, P. (2008). Response shift and Quality of Life (QOL): A priming study. *Psychology & Health*, *23*, 63. doi:10.1007/s11136-010-9737-y.



- Bernhard, J., Lowy, A., Maibach, R., & Hurny, C. (2001). Response shift in the perception of health for utility evaluation. *An explorative investigation. European Journal of Cancer, 37*, 1729–1735.
- Boucekine, M., Loundou, A., Baumstarck, K., Minaya-Flores, P., Pelletier, J., Ghattas, B., & Auquier, P. (2013). Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: A cohort study. *BMC Medical Research Methodology, 13*, 20. doi:10.1186/1471-2288-13-20.
- Bray, J. H., Maxwell, S. E., & Howard, G. S. (1984). Methods of analysis with response-shift bias. *Educational and Psychological Measurement, 44*, 781–804.
- Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Boca Raton, FL: Chapman & Hall.
- Cohen, S. R., Mount, B. M., Tomas, J. J., & Mount, L. F. (1996). Existential well-being is an important determinant of quality of life: Evidence from the McGill Quality of Life Questionnaire. *Cancer, 77*, 576–586.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi:10.1037/h0040957.
- Department of Health. (2009). *Guidance on the routine collection of Patient Reported Outcomes Measures (PROMs)*. Retrieved from [http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH\\_092647](http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_092647).
- Fayers, P., & Machin, D. (2007). *Quality of life: The assessment, analysis and interpretation of patient-reported outcomes*. Chichester, UK: John Wiley & Sons.
- Guilleux, A., Blanchin, M., Vanier, A., Guillemin, F., Falissard, B., Schwartz, C. E., et al. (2015). ResONse Shift ALgorithm in Item response theory (ROSALI) for response shift detection with missing data in longitudinal patient-reported outcome studies. *Quality of Life Research, 24*, 553–564. doi:10.1007/s11136-014-0876-4.
- Hickey, A. M., Bury, G., O’Boyle, C. A., Bradley, F., O’Kelly, F. D., & Shannon, W. (1996). A new short form individual quality of life measure (SEIQoL-DW): Application in a cohort of individuals with HIV/AIDS. *British Medical Journal, 313*, 29–33.
- Huberty, C. J., & Wisenbaker, J. M. (1992). Variable importance in multivariate group comparisons. *Journal of Educational and Behavioral Statistics, 17*, 75–91.
- Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123*, 207–215.
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219–230. doi:10.1007/s11205-011-9843-4.
- Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 3–19). Washington, DC: American Psychological Association.
- Jobe, J. B. (2003). Cognitive psychology and self-reports: Models and methods. *Quality of Life Research, 12*, 219–227.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. doi:10.1111/jedm.12000.
- Kim, S., & Camilli, G. (2014). An item response theory approach to longitudinal analysis with application to summer setback in preschool language/literacy. *Large-scale Assessments in Education, 2*, 1–17. doi:10.1186/2196-0739-2-1.
- King-Kallimanis, B. L., Oort, F. J., & Garst, G. J. A. (2010). Using structural equation modelling to detect measurement bias and response shift in longitudinal data. *AStA Advances in Statistical Analysis, 94*, 139–156. doi:10.1007/s10182-010-0129-y.



- King-Kallimanis, B. L., Oort, F. J., Nolte, S., Schwartz, C. E., & Sprangers, M. A. (2011). Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients. *Quality of Life Research*, *20*, 1527–1540. doi:[10.1007/s11136-010-9844-9](https://doi.org/10.1007/s11136-010-9844-9).
- King-Kallimanis, B. L., Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Structural equation modeling of health-related quality-of-life data illustrates the measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, *62*, 1157–1164. doi:[10.1016/j.jclinepi.2009.04.004](https://doi.org/10.1016/j.jclinepi.2009.04.004).
- Li, Y., & Rapkin, B. (2009). Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *Journal of Clinical Epidemiology*, *62*, 1138–1147. doi:[10.1016/j.jclinepi.2009.03.021](https://doi.org/10.1016/j.jclinepi.2009.03.021).
- Lix, L. M., Chan, E. K. H., Sawatzky, R., Sajobi, T. T., Liu, J., Hopman, W., & Mayo, N. (2015). Response shift and disease activity in inflammatory bowel disease. *Quality of Life Research*, *25*, 1751–1760. doi:[10.1007/s11136-015-1188-z](https://doi.org/10.1007/s11136-015-1188-z).
- Lix, L. M., Sajobi, T. T., Sawatzky, R., Liu, J., Mayo, N. E., Huang, Y., et al. (2013). Relative importance measures for reprioritization response shift. *Quality of Life Research*, *22*, 695–703. doi:[10.1007/s11136-012-0198-3](https://doi.org/10.1007/s11136-012-0198-3).
- Lowy, A., & Bernhard, J. (2004). Quantitative assessment of changes in patients' constructs of quality of life: An application of multilevel models. *Quality of Life Research*, *13*, 1177–1185.
- Martin, F., Camfield, L., Rodham, K., Kliempt, P., & Ruta, D. (2007). Twelve years' experience with the Patient Generated Index (PGI) of quality of life: A graded structured review. *Quality of Life Research*, *16*, 705–715. doi:[10.1007/s11136-006-9152-6](https://doi.org/10.1007/s11136-006-9152-6).
- Mayo, N. E., Scott, S. C., Dendukuri, N., Ahmed, S., & Wood-Dauphinee, S. (2008). Identifying response shift statistically at the individual level. *Quality of Life Research*, *17*, 627–639. doi:[10.1007/s11136-008-9329-2](https://doi.org/10.1007/s11136-008-9329-2).
- McClimans, L., Bickenbach, J., Westerman, M., Carlson, L., Wasserman, D., & Schwartz, C. (2013). Philosophical perspectives on response shift. *Quality of Life Research*, *22*, 1871–1878. doi:[10.1007/s11136-012-0300-x](https://doi.org/10.1007/s11136-012-0300-x).
- Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*, 279–300. doi:[10.1207/s15327574ijt0503\\_6](https://doi.org/10.1207/s15327574ijt0503_6).
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan Publishing Co Inc.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
- Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*, 35–44.
- Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*, 5–9. doi:[10.1111/j.1750-8606.2009.00109.x](https://doi.org/10.1111/j.1750-8606.2009.00109.x).
- Neuman, H. B., Park, J., Fuzesi, S., & Temple, L. K. (2012). Rectal cancer patients' quality of life with a temporary stoma: Shifting perspectives. *Diseases of the Colon and Rectum*, *55*, 1117–1124. doi:[10.1097/DCR.0b013e3182686213](https://doi.org/10.1097/DCR.0b013e3182686213).
- Nolte, S., Elsworth, G. R., Sinclair, A. J., & Osborne, R. H. (2009). Tests of measurement invariance failed to support the application of the “then-test”. *Journal of Clinical Epidemiology*, *62*, 1173–1180. doi:[10.1016/j.jclinepi.2009.01.021](https://doi.org/10.1016/j.jclinepi.2009.01.021).
- Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*, 587–598.
- Oort, F. J., Visser, M. R., & Sprangers, M. A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research*, *14*, 599–609.
- Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, *62*, 1126–1137. doi:[10.1016/j.jclinepi.2009.03.013](https://doi.org/10.1016/j.jclinepi.2009.03.013).

- Rapkin, B. D., & Schwartz, C. E. (2004). Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. *Health and Quality of Life Outcomes*, 2, 14. doi:[10.1186/1477-7525-2-14](https://doi.org/10.1186/1477-7525-2-14).
- Rapkin, B. D., & Schwartz, C. E. (2016). Distilling the essence of appraisal: A mixed methods study of people with multiple sclerosis. *Quality of Life Research*, 25, 793–805.
- Ruta, D. A., Garratt, A. M., Leng, M., Russell, I. T., & MacDonald, L. M. (1994). A new approach to the measurement of quality of life. The Patient-Generated Index. *Medical Care*, 32, 1109–1126.
- Sajobi, T. T., Fiest, K. M., & Wiebe, S. (2014). Changes in quality of life after epilepsy surgery: The role of reprioritization response shift. *Epilepsia*, 55, 1331–1338. doi:[10.1111/epi.12697](https://doi.org/10.1111/epi.12697).
- Sajobi, T. T., Lix, L. M., Clara, I., Walker, J., Graff, L. A., Rawsthorne, P., et al. (2012). Measures of relative importance for health-related quality of life. *Quality of Life Research*, 21, 1–11. doi:[10.1007/s11136-011-9914-7](https://doi.org/10.1007/s11136-011-9914-7).
- Sajobi, T. T., Lix, L. M., Singh, G., Lowerison, M., Engbers, J., & Mayo, N. E. (2015). Identifying reprioritization response shift in a stroke caregiver population: A comparison of missing data methods. *Quality of Life Research*, 24, 529–540. doi:[10.1007/s11136-014-0824-3](https://doi.org/10.1007/s11136-014-0824-3).
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.
- Sawatzky, R., Chan, E. K. H., Zumbo, B. D., Bingham, C. O., Gardner, W., Jutai, J., . . . Lix, L. M. (in press). Challenges and opportunities in patient-reported outcomes validation. *Journal of Clinical Epidemiology*. doi:[10.1016/j.jclinepi.2016.12.002](https://doi.org/10.1016/j.jclinepi.2016.12.002)
- Sawatzky, R., Gadermann, A., Ratner, P. A., Zumbo, B. D., & Lix, L. M. (2012). Identifying individuals with inflammatory bowel disease who experienced response shift: A latent class analysis. *Quality of Life Research*, 21, 33.
- Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research*, 17, 343–358.
- Schmitt, N., Pulakos, E., & Lieblein, A. (1984). A comparison of three techniques to assess group-level beta and gamma change. *Applied Psychological Measurement*, 8, 249–260.
- Schwartz, C. E., Ahmed, S., Sawatzky, R., Sajobi, T., Mayo, N., Finkelstein, J., et al. (2013). Guidelines for secondary analysis in search of response shift. *Quality of Life Research*, 22, 2663–2673. doi:[10.1007/s11136-013-0402-0](https://doi.org/10.1007/s11136-013-0402-0).
- Schwartz, C. E., Bode, R., Repucci, N., Becker, J., Sprangers, M. A., & Fayers, P. M. (2006). The clinical significance of adaptation to changing health: A meta-analysis of response shift. *Quality of Life Research*, 15, 1533–1550. doi:[10.1007/s11136-006-0025-9](https://doi.org/10.1007/s11136-006-0025-9).
- Schwartz, C. E., & Rapkin, B. A. (2012). Understanding appraisal processes underlying the thentest: A mixed methods investigation. *Quality of Life Research*, 21, 381–388. doi:[10.1007/s11136-011-0023-4](https://doi.org/10.1007/s11136-011-0023-4).
- Schwartz, C. E., Sajobi, T. T., Verdam, M. G., Seville, V., Lix, L. M., Guilleux, A., & Sprangers, M. A. (2015). Method variation in the impact of missing data on response shift detection. *Quality of Life Research*, 24, 521–528. doi:[10.1007/s11136-014-0746-0](https://doi.org/10.1007/s11136-014-0746-0).
- Schwartz, C. E., & Sprangers, M. A. (2009). Reflections on genes and sustainable change: Toward a trait and state conceptualization of response shift. *Journal of Clinical Epidemiology*, 62, 1118–1123. doi:[10.1016/j.jclinepi.2009.02.008](https://doi.org/10.1016/j.jclinepi.2009.02.008).
- Schwartz, C. E., & Sprangers, M. A. (2010). Guidelines for improving the stringency of response shift research using the thentest. *Quality of Life Research*, 19, 455–464. doi:[10.1007/s11136-010-9585-9](https://doi.org/10.1007/s11136-010-9585-9).
- Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science and Medicine*, 48, 1507–1515.
- Thomas, D. R. (1992). Interpreting discriminant functions: A data analytic approach. *Multivariate Behavioral Research*, 27, 335–362. doi:[10.1207/s15327906mbr2703\\_3](https://doi.org/10.1207/s15327906mbr2703_3).
- Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45, 253–275.
- Thomas, D. R., Zhu, P., Zumbo, B. D., & Dutta, S. (2008). On measuring the relative importance of explanatory variables in a logistic regression. *Journal of Modern Applied Statistical Methods*, 7, 21–38.

- Thomas, D. R., & Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics, 21*, 110–130. doi:[10.3102/10769986021002110](https://doi.org/10.3102/10769986021002110).
- Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge, U.K.: Cambridge University Press.
- Verdam, M. G., Oort, F. J., & Sprangers, M. A. G. (2016). Using structural equation modeling to detect response shifts and true change in discrete variables: An application to the items of the SF-36. *Quality of Life Research, 25*, 1361–1383. doi:[10.1007/s11136-015-1195-0](https://doi.org/10.1007/s11136-015-1195-0).
- Verdam, M. G., Oort, F. J., van der Linden, Y. M., & Sprangers, M. A. (2015). Taking into account the impact of attrition on the assessment of response shift and true change: A multigroup structural equation modeling approach. *Quality of Life Research, 24*, 541–551. doi:[10.1007/s11136-014-0829-y](https://doi.org/10.1007/s11136-014-0829-y).
- Westerman, M. J., Hak, T., Sprangers, M. A., Groen, H. J., van der Wal, G., & The, A. M. (2008). Listen to their answers! Response behaviour in the measurement of physical and role functioning. *Quality of Life Research, 17*, 549–558. doi:[10.1007/s11136-008-9333-6](https://doi.org/10.1007/s11136-008-9333-6).
- Zumbo, B. D. (2007a). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26: Psychometrics, pp. 45–79). Amsterdam, the Netherlands: Elsevier Science.
- Zumbo, B. D. (2007b). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223–233.
- 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, NC: Information Age Publishing.
- Zumbo, B. D., & Chan, E. K. H. (2014). Reflections on validation practices in the social, behavioral, and health sciences. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (Vol. 54, pp. 321–327). New York, NY: Springer International Publishing.

# Chapter 15

## Validating a Distractor-Driven Geometry Test Using a Generalized Diagnostic Classification Model

Benjamin R. Shear and Louis A. Roussos

### Introduction

This chapter explores the use of a recently developed psychometric model (DiBello, Henson, & Stout, 2015) intended for selected response tests with option-based scoring. The model has the potential to bridge two promising but largely independent areas in the development of formative assessments: distractor-driven assessments of student misconceptions (e.g., Hestenes, Wells, & Swackhamer, 1992; Masters, 2012b; Russell, O'Dwyer, & Miranda, 2009; Sadler, 1998) and diagnostic classification models (DCM; Rupp, Templin, & Henson, 2010). Toward this end, the model is used to contribute additional validity evidence supporting the intended interpretations and uses of scores on a middle school geometry test, as advocated by professional testing standards (AERA, APA, & NCME, 2014). Specifically, a DCM is used to evaluate whether student response patterns on a middle school mathematics test developed as part of the Diagnostic Geometry Assessment (Masters, 2010) project are consistent with response processes test-takers are hypothesized to be using.

---

B.R. Shear (✉)

School of Education, University of Colorado Boulder, 249 UCB, Boulder, CO 80309, USA  
e-mail: [benjamin.shear@colorado.edu](mailto:benjamin.shear@colorado.edu)

L.A. Roussos

Measured Progress, 100 Education Way, Dover, NH 03820, USA  
e-mail: [roussos.louis@measuredprogress.org](mailto:roussos.louis@measuredprogress.org)

© 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\_15

277

## ***Validity and Validation***

The evaluation of test score interpretations and uses falls under the heading of “validity,” which is considered “the most fundamental consideration in developing tests and evaluating tests” (AERA et al., 2014, p. 11). There is ongoing debate among scholars about how best to define the term “validity” (e.g., Lissitz, 2009). The *Standards for Educational and Psychological Testing* (henceforth the *Standards*; AERA et al., 2014), building on the seminal work of Messick (1989), provide the following definitions for validity and validation to guide test developers and users in the field of psychometrics:

Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests... The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself. (AERA et al., 2014, p. 11)

Test validity thus provides a conceptual framework for evaluating the uses and interpretations of test scores, while test validation encompasses the methods and methodologies used to carry out the evaluation of test score interpretations and uses (Borsboom, Mellenbergh, & van Heerden, 2004; Zumbo, 2007).

The *Standards* do not provide a specific set of steps that can be followed during the process of validation. Instead, the *Standards* describe five potential sources (or “types”) of validity evidence that might be gathered in the process of validation. These include evidence based on: (a) the relations of scores to other variables, (b) the internal structure of test-taker responses, (c) response processes, (d) test content, and (e) consequences of test use and interpretation. Although these categories are not mutually exclusive, they encompass a broad range of the types of studies or analyses that are typically carried out in the process of validation. Current discussions of validity theory (e.g., Kane, 2013; Messick, 1989; Sireci, 2009; Zumbo, 2007) consistently highlight a number of key points including: (a) validity is a matter of degree, not all or none; (b) test score interpretations and uses are validated, not tests themselves; (c) a judgment of degree of validity should draw on a wide range of theoretical and empirical evidence. Although there are some alternative definitions of validity in the literature (e.g., Borsboom et al., 2004; Lissitz & Samuelsen, 2007), we adopt the definitions of validity and validation as described in the *Standards*.

## ***Validating Distractor-Driven Tests***

Distractor-driven tests, such as the Force Concept Inventory (FCI; Hestenes et al., 1992) or tests using ordered multiple choice items (Briggs, Alonzo, Schwab, & Wilson, 2006), are selected-response tests intended to provide timely and useful diagnostic information about student understanding, including identifying students that may be reasoning with systematic misconceptions. These selected-response

tests include carefully chosen “incorrect” response options (henceforth “distractors”) corresponding to common student mistakes and misconceptions. Scoring of these tests takes into account not only whether an item was answered correctly or not, but also which incorrect response options were selected across items. These tests use “option-based” scoring to create more complex scores, intended to provide more detailed and useful instructional feedback to teachers than traditional selected-response tests that generally only indicate the number of correctly answered questions. At the same time, the selected-response format makes distractor-driven tests timely and efficient for classroom use.

Interpreting scores on distractor-driven tests involves stronger claims about the response processes used by test-takers than many other test scores. With a mathematics achievement test used to determine student proficiency, for example, test users may be more interested in the number of questions students answer correctly rather than the strategies used to answer them. With many distractor-driven tests, however, the aim is to make inferences or claims about how students are reasoning to inform subsequent instruction. Moreover, distractor-driven tests are often intended to measure multidimensional attributes that include both problematic and desirable aspects of student reasoning. Although evidence based on response processes is rarely presented in practice (e.g., Zumbo & Chan, 2014), such evidence is vital for distractor-driven tests. DCMs provide a promising way to provide such evidence for distractor-driven tests.

DCMs are part of a larger family of psychometric and statistical models that can be used to classify test-takers into latent profiles based on observed responses (Rupp et al., 2010). In contrast to more common unidimensional psychometric models primarily intended for scaling respondents along a continuous latent dimension, DCMs are intended to classify respondents into two or more discrete latent classes. Use of DCMs begins by positing a set of skills or attributes that are measured by items or tasks on a test. Different skill or attribute profiles are defined by having different combinations of these skills or attributes. Statistical models are then used to classify test-takers, based on their item responses, into one of the different attribute profiles. Related literature often falls under the heading of “cognitive diagnostic assessment” (Leighton & Gierl, 2007a), “cognitive skills models” (Roussos, Templin, & Henson, 2007), and various other names (Rupp et al., p. 3).

There are several potential benefits to using DCMs in the development and validation of distractor-driven tests. First, DCMs provide a statistical framework for operationalizing and testing cognitive theories about examinee response processes that matches the context of distractor-driven tests. This includes recently developed models that can account for nominal item responses and multidimensionality (e.g., DiBello et al., 2015). Second, DCMs provide model-based classifications of examinees that can be used to evaluate proposed score interpretations and that are likely to be more reliable than classifications based on item response theory models with continuous latent variables (e.g., Templin & Bradshaw, 2013). Third, DCMs provide item-level information that can be used to inform item revision or creation, including facilitating the building of test forms with consistent properties (Roussos, DiBello, Henson, Jang, & Templin, 2010).

This chapter illustrates how the generalized diagnostic classification model for multiple-choice option-based scoring (GDCM-MC; DiBello et al., 2015) can be used to provide validity evidence related to response processes for a geometry test using option-based scoring. Specifically, the test of interest was developed as part of the Diagnostic Geometry Assessment project conducted by Masters (2010, 2012a). This test provided a particularly attractive application in that it was developed independently of the psychometric model while also embracing many salient features that the model was designed to be sensitive to: multidimensional constructs, items with options systematically designed in accordance with the principles underlying the model, and a diagnostic classification purpose aligned with the purpose of the psychometric model. To be most useful for informing and improving instruction in a formative assessment system, DCMs need to be used within a framework that includes theories about cognitive processes and carefully constructed tests (Roussos et al., 2010), and this example seems to provide a situation that is remarkably well aligned with this appeal. Thus, the GDCM-MC model is applied to a single test form to address the following research questions: (1) Do patterns in students' responses support the proposed diagnostic interpretations based on a one- or a two-dimensional misconception? (2) Can the validity of the original scoring model be improved using information from the GDCM-MC results?

## **Background on Shape Properties Test**

The Diagnostic Geometry Assessment Project (Masters, 2010, 2012a, 2012b, 2014) used research in cognition and learning to develop a set of three selected-response tests for use as formative classroom assessments (Black & Wiliam, 1998) in middle school geometry. The test development process began by identifying three "misconceptions" that students often hold when learning basic concepts in geometry. The test developers followed prior researchers (e.g., Smith, diSessa, & Roschelle, 1993) in viewing misconceptions not simply as incorrect understandings, but as an important developmental stage of students' knowledge construction process. In this way they are similar to the notion of "facets" of student thinking (e.g., Minstrell, 2000). The goal of the Diagnostic Geometry Assessment project was to develop tests that could inform teachers' subsequent instruction by providing information about their students' misconceptions. The test developers identified three commonly held misconceptions in the research literature and then developed one test and accompanying instructional materials targeting each misconception. This chapter focuses on the "shape properties" test.



### ***Development of the Shape Properties Test***

The shape properties test is intended to help teachers determine whether students are reasoning with a concept image misconception (Masters, 2010; Vinner & Hershkowitz, 1980). A “concept image” is a visual example of a geometric concept that students may use when learning about geometric shapes. This can be contrasted with a “concept definition,” which is a formal definition of a geometric concept. When students encounter geometric shapes and need to classify them, they may base their classification on whether the presented shape is similar to one of their concept images rather than whether it satisfies the formal concept definition. While this reasoning is partially correct and can result in the correct classification of many shapes, it can also lead to systematic errors for other shapes. Knowing whether students might be relying primarily on concept images, rather than concept definitions, could provide helpful information for teachers.

The shape properties test presents students with a series of selected-response questions asking them to classify sets of rectangles and parallelograms. These items were developed through an iterative process that began by having students answer open-ended questions about simple geometric shapes, often requiring them to classify the shapes being presented (Masters, 2010). The test developers then identified common responses that represented either (a) correct answers, (b) incorrect answers that showed evidence of reasoning with the concept image misconception, or (c) incorrect answers resulting from other errors. These responses were used to create a single correct response option and a mix of incorrect options (distractors) that may or may not represent evidence of reasoning with the misconception. These selected-response items then underwent further pilot testing and analysis, including cognitive labs and classroom administrations, to select a final set of 12 items. The research team also developed curriculum materials and activities that teachers could use based on the results of administering the shape properties and other tests.

### ***Scoring the Shape Properties Test***

The 12 items on the shape properties test include five standard multiple-choice questions (each with four options) and seven questions comprised of between three and eight sets of binary selected-response options that relate to a single item stem. There are four possible responses for the standard multiple-choice items (students can select any one of the four options) and there are anywhere from  $2^3$  to  $2^8$  possible response patterns for the binary sets of choices items. Any given response pattern for an item can be coded as either (a) correct, (b) incorrect and consistent with the concept image misconception (a “misconception response”), or (c) incorrect but not consistent with the concept image misconception (an “incorrect response”). There is only one correct response pattern for each item, but there can be multiple incorrect and multiple misconception response patterns for an item.

Answer each question below about **all** rectangles. Choose **Yes** or **No** for each option.

Yes  No Do all rectangles have 4 sides?


Yes  No Do all rectangles have 2 sides that are longer than the other 2 sides?

Yes  No Do all rectangles have 4 sides of equal length?

Yes  No Do all rectangles have 2 pairs of parallel sides?

Yes  No Do all rectangles have 4 right angles?

What would you have to change about this rectangle for it to become a parallelogram?



A Tilt the shorter sides to the right to create acute and obtuse angles.

B There is no way to change this rectangle into a parallelogram.

C Rotate the rectangle 90 degrees.

D Nothing. The rectangle is already a parallelogram.

Fig. 15.1 Sample shape properties test items with answer keys

Figure 15.1 presents two sample items from the shape properties test; the top item displays an example of the binary pairs of response options with five pairs of yes/no questions, and the lower item displays an example of a standard 4-option multiple-choice question. The items as presented in Fig. 15.1 include the scoring key. In the top item, the correct response pattern (yes-no-no-yes-yes) is represented by the filled response options. Any response pattern that includes the responses enclosed in dashed boxes (i.e., any response of the form yes-yes-X-yes-X where “X” can be either a yes or a no) would be considered a misconception response. Any

other response would be considered an incorrect response. For the lower item, selecting option A (the first option) is a misconception response, selecting options B or C are incorrect responses, and selecting option D (the last option) is a correct response.

In prior studies (e.g., Masters, 2014) the shape properties test was scored twice, using two different scoring keys, in a process similar to using multiple scoring evaluators (Luecht, 2007). First, the number of correct responses was used to calculate an “ability score” for each student. Second, the number of misconception responses was used to calculate a “misconception score.” Then, based on research with tests used to diagnose student misconceptions in algebra (Russell et al., 2009), in addition to analysis of data collected from the cognitive labs and pilot studies, the researchers classified students answering 75% or more of the items correctly (9 out of 12 or more) as having mastered the desired understanding, and students providing misconception responses to 35% or more of the items (5 out of 12 or more) as reasoning with the target misconception. All other students were classified as “mistakers.” Using these rules, students can be classified as “knowers,” “misconceivers,” or “mistakers.” With this scoring, the three classifications are mutually exclusive.

### *Validity Questions About the Shape Properties Test*

When responses to the shape properties test were evaluated empirically in a pilot sample, preliminary analyses (including principal components analyses and use of a Rasch item response theory model) suggested the misconception might comprise two distinct dimensions, depending upon whether the items make reference to parallelograms (Masters, 2010). Further analysis of the items revealed that five of the items (referred to here as “items 1–5”) involve identifying sub-categories of rectangles without reference to parallelograms, while the other seven items (referred to here as “items 6–12”) involve parallelograms explicitly. To explore the possibility of a multidimensional misconception, the researchers applied the scoring and classification rules (e.g., 35% or more misconception responses indicates a misconceiver) to the item types separately (Masters, 2010, Table 38). The two different sets of items were found to classify students differently suggesting the test might measure two distinct aspects of the misconception.

Figure 15.1 illustrates the difference between the misconceptions measured by the two sets of items. The top item in Fig. 15.1 is from items 1–5 (not referencing parallelograms) and the lower item is from items 6–12. Both items measure whether students understand and can correctly classify examples of parallelograms based on their properties, which is the desired understanding. If the test measures a single misconception, then both items would provide information about whether students are reasoning with that misconception. The prior results suggest defining two separate aspects of the misconception may be more appropriate. The two misconceptions could be defined as: (1) students reason with a misconception about squares

and rectangles based on an image of rectangles with long and short sides, and (2) students reason with a misconception about parallelograms based on an image that parallelograms always have pairs of long and short sides that are “tilted.”

Thus, while there is a growing body of validity evidence to support the use and interpretation of scores on the shape properties test for formative purposes (e.g., Masters, 2012a), to date there have not been analyses of the internal structure of the test that adequately incorporate the hypothesized response processes and misconception reasoning. As described above, DCMs provide a natural way to provide this type of validity evidence. First, DCMs can be used to evaluate whether student response patterns correspond to the types of response patterns that would be expected if the test is measuring the intended misconceptions. For example, models with one- and two-dimensional misconceptions could be directly compared. Second, the model-based classifications of students can be used to evaluate the proposed observed-score classification rules described above and gain a better understanding of typical student response profiles.

## Statistical Models

Most DCMs are intended to analyze binary item response data where items are scored correct/incorrect. Two common models are the deterministic inputs noisy and gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001; Rupp et al., 2010) and the reduced reparametrized unified model (R-RUM; Roussos, DiBello, et al., 2007; Rupp et al., 2010). At least three DCMs for modeling nominal response options have appeared recently in the published literature (Bradshaw & Templin, 2014; de la Torre, 2009a; DiBello et al., 2015). The MC-DINA model (de la Torre, 2009a) entails overly restrictive constraints for the present analyses. The SICM model (Bradshaw & Templin, 2014) is intended to scale test-takers along a continuous dimension while also identifying misconception-based reasoning, hence entailing the estimation of additional parameters not required for the present use. This study thus uses the GDCM-MC model (DiBello et al., 2015), which allows for a flexible range of item response functions and includes an additional component quantifying random guessing behaviors. To introduce the GDCM-MC modeling framework, we first describe two more basic binary item response DCMs.

## *Overview of Diagnostic Classification Models*

Use of DCMs begins by defining a set of discrete latent attributes that a test is intended to measure or provide information about.<sup>1</sup> Second, a “Q matrix” (Tatsuoka, 1983) indicating which items measure which attributes is constructed. Third, a

---

<sup>1</sup>The generic term “attribute” will be used throughout to refer to the constructs the test is intended to measure or to provide information about.

cognitive link function describing how the different attributes influence item responses is defined. Once these three pieces are defined, statistical models are used to classify examinees into different attribute profiles based on their observed responses to the test items. An examinee’s true attribute profile is never directly observed, but is instead inferred based on his or her item responses.

In the case of the shape properties test, the attributes include the desired understanding and the one (or two) targeted misconceptions. For a test that measures  $K$  attributes, one can then define a  $1 \times K$  attribute vector,  $\alpha_j$ , for examinee  $j$ , where  $\alpha_{jk} = 1$  if examinee  $j$  possesses attribute  $k$  and  $\alpha_{jk} = 0$  otherwise. In the case of the shape properties test, a student reasoning with one of the misconceptions “possesses” the attribute representing that misconception in the model. Each possible  $\alpha$  vector defines an “attribute profile.”

Next, suppose there are  $I$  items on a test, where each item requires a (non-empty) subset of the attributes to answer correctly (or to endorse), and each attribute is required by at least one item on the test. The  $Q$  matrix is an  $I \times K$  matrix with a column for each attribute and a row for each item that has entries  $q_{ik} = 1$  if item  $i$  requires attribute  $k$  and  $q_{ik} = 0$  if item  $i$  does not require attribute  $k$ . The  $Q$  matrix is specified a priori, and is akin to defining a factor pattern matrix in a confirmatory factor analysis (Templin, 2016).

The cognitive link function specifies how attributes relate to item responses. In the DINA model, for example, the response function representing the probability of a correct response for examinee  $j$  to item  $i$  is:

$$P(X_{ij} = 1) = (1 - s_i)^{\xi_{ij}} g_i^{1 - \xi_{ij}}, \tag{15.1}$$

where<sup>2</sup>

$$\xi_{ij} = \prod_{k=1}^K \alpha_{jk}^{q_{ik}}, \tag{15.2}$$

which implies that

$$\xi_{ij} = \begin{cases} 1 & \text{if } \alpha_{jk} \geq q_{ik} \forall k, \\ 0 & \text{otherwise} \end{cases} \tag{15.3}$$

Additional constraints are usually imposed, for example that  $(1 - s_i) > g_i$ . Here  $\xi_{ij} = 1$  if and only if examinee  $j$  possesses all attributes required for item  $i$  (elements for which  $q_{ik} = 1$ ), and  $\xi_{ij} = 0$  otherwise. This model postulates that if an examinee possesses all required attributes for an item, they will answer it correctly with

---

<sup>2</sup>This follows common practice (e.g., de la Torre, 2009b; Rupp, Templin, & Henson, 2010) in defining the model for notational convenience. Note, however, that this notation can sometimes imply the operation  $0^0$  which is mathematically indeterminate. Here, one can define  $0^0 = 1$  for convenience, to imply that if a test-taker does not possess a non-required attribute, it will not reduce their likelihood of answering the item correctly.

probability  $(1 - s_i)$ ; if an examinee lacks one or more required attributes, they will answer it correctly with probability  $g_i$ . While different attribute profiles may have different predicted likelihoods of success across multiple items, many profiles can have the same predicted probability of success on a single item (based on whether the profile includes all required attributes or not).

The R-RUM model increases the differentiation in predicted probabilities across attribute profiles by specifying the following response function for examinee  $j$  to item  $i$ :

$$P(X_{ij} = 1) = \pi_i \prod_{k=1}^{K_i} r_{ik}^{(1-\alpha_{jk}) q_{ik}} \quad (15.4)$$

where  $0 < \pi_i < 1$  and  $0 < r_{ik} < 1$ . With this link function, an examinee who possesses all attributes required for an item has probability  $\pi_i$  of answering the item correctly. If attribute  $k$  is required for an item but the examinee does not possess the attribute, the likelihood of success is reduced by the multiplicative factor  $r_{ik}$ , that is  $\pi_{ij} = \pi_i r_{ik}$ . With the R-RUM there can be many more predicted probabilities of success for each item, depending upon which attributes an examinee does or does not possess. When item  $i$  requires  $K_i$  attributes, then when  $K_i > 1$  the model requires the estimation of significantly more parameters than the DINA model, as there are  $K_i + 1$  parameters per item.

### *The Generalized Diagnostic Classification Model*

The GDCM-MC model (DiBello et al., 2015) extends these and other binary DCMs for the analysis of nominal multiple-choice items in two important ways. First, the GDCM-MC expands the Q matrix to include a row for each option of each item, rather than having only a single row for each item (the Q matrix retains a column for each attribute). Second, the Q matrix entries can be 0, 1, or N, rather than only 0 or 1. These extensions allow the GDCM-MC to be more flexible and incorporate problematic attributes (e.g., misconceptions), but it also introduces some complications. These are described more fully in the next section.

The GDCM-MC Q matrix contains entries  $q_{ihk}$ , where there are  $H_i$  options for item  $i$ . In the new coding scheme, a  $q_{ihk} = 1$  indicates that possessing attribute  $k$  increases the likelihood of choosing option  $h$ , and thus selecting that response option provides evidence that the test-taker possesses attribute  $k$ . Meanwhile  $q_{ihk} = 0$  indicates that possessing attribute  $k$  decreases the likelihood of choosing option  $h$ , and thus selecting that response option provides evidence that the test-taker does not possess attribute  $k$ . This contrasts with dichotomous DCMs, where a 0 entry often indicates that possession of a particular attribute does not impact the likelihood of answering an item correctly, and thus a response on that item does not provide evidence about possessing attribute  $k$ . Lastly,  $q_{ihk} = N$  has a similar meaning to a 0 in

a binary DCM and indicates that possessing attribute  $k$  does not directly impact the likelihood of selecting option  $h$ , thus providing a lack of evidence about whether the test-taker possesses attribute  $k$ . There are important subtleties to these interpretations of the 0, 1, and  $N$  terms that become more complex to interpret, as described below.

We use a form of the GDCM-MC that extends the R-RUM model. Once the  $Q$  matrix has been constructed, the following cognitive link function is used to model the probability that examinee  $j$  selects response option  $h$  to item  $i$  based on the profile vector  $\alpha_j$ :

$$P_i(X_{ij} = h | \alpha_j) = \begin{cases} F_{ih}(\alpha_j) + (1 - S_{i,\alpha_j}) \frac{1}{H_i} & \text{if } S_{i,\alpha_j} < 1 \\ \frac{F_{ih}(\alpha_j)}{S_{i,\alpha_j}} & \text{if } S_{i,\alpha_j} \geq 1 \end{cases} \quad (15.5)$$

The core of the model is  $F_{ih}(\alpha_j)$ , an extension of the R-RUM link function in Eq. 15.4 that contains  $\pi_{ih}$  and  $r_{ihk}$  parameters specific to each option of each item and takes an attribute profile as input:

$$F_{ih}(\alpha_j) = \pi_{ih} \prod_{k \text{ such that } q_{ihk} \neq N} r_{ihk}^{|q_{ihk} - \alpha_{jk}|} \quad (15.6)$$

As with the R-RUM, the response function contains a series of  $\pi$  and  $r$  parameters that come into play under different circumstances. Here, the  $|q_{ihk} - \alpha_{jk}|$  term means that the penalty parameters are activated each time an element of the examinee’s profile vector,  $\alpha_{jk}$ , mismatches a non- $N$  element  $q_{ihk}$  of an option’s  $Q$  matrix row. To ensure that

$$\sum_{h=1}^{H_i} P_i(X_{ij} = h | \alpha_j) = 1.0 \quad (15.7)$$

(i.e., that the option function values sum to 1 and represent probabilities), the  $F_{ih}(\alpha_j)$  values are re-scaled by their sum:

$$S_{i,\alpha_j} = \sum_{h=1}^{H_i} F_{ih}(\alpha_j), \quad (15.8)$$

if  $S_{i,\alpha_j} \geq 1$ . If  $S_{i,\alpha_j} < 1$  then the  $F_{ih}(\alpha_j)$  values are treated as probabilities and the model assumes that students will randomly select one of the response options with probability  $(1 - S_{i,\alpha_j})$  to ensure that Eq. 15.7 is satisfied. The term  $(1 - S_{i,\alpha_j})$ , included only when  $S_{i,\alpha_j} < 1$ , represents the probability of random guessing on an item and can be a substantively interesting parameter to interpret. The  $S_{i,\alpha_j}$  parameter also introduces a complex dependence among the response option probabilities discussed further below.



## Methods

### *Q Matrix Construction*

Q matrix specification is a critical step in the application of DCMs because it operationalizes the hypothesized cognitive model (Madison & Bradshaw, 2015; Rupp & Templin, 2008). In earlier work (DiBello, Henson, Stout, & Roussos, 2014) researchers used the GDCM-MC model to analyze student responses to a different Diagnostic Geometry Assessment project test, the “geometric measurement” test. The Q matrix construction process in that study involved re-defining the target attributes and the links between these attributes and the response options, relative to the initial scoring keys proposed by the test developers. This study uses a different Q matrix construction process. Specifically, a confirmatory approach is used to specify Q matrices that seek to operationalize the attributes and response processes the test developers originally hypothesized to underlie student responses to the shape properties test, directly utilizing the original scoring key. Four Q matrices were compared – one operationalizing a response process with a single misconception attribute and three operationalizing response processes with two misconceptions. Three versions of the two-dimensional Q matrices were used because it became clear that there was more than one way to operationalize the response processes entailed by the original scoring key within the extended Q matrix framework.

The first Q matrix assumes there is a single skill attribute measured by the test (correct understanding of classifying geometric shapes) and a single misconception attribute representing reasoning with the concept image misconception. For this Q matrix, every item has the same Q matrix format; the format is shown in Table 15.1. The attribute vectors are organized as  $\alpha_j = [I_S, I_M]$ , where  $I_S$  and  $I_M$  are 0/1 indicator variables indicating possession of the skill and misconception attributes, respectively. The first row, for example, indicates that students selecting an incorrect response option that is not consistent with the misconception are most likely to possess neither the desired understanding nor the target misconception, because the attribute profile matching this row would be  $\alpha_j = [0, 0]$ . The second row indicates that a student selecting an incorrect response option that is consistent with the target misconception is most likely to possess the target misconception, and have attribute profile  $\alpha_j = [0, 1]$ . A student selecting the correct option is expected to have the desired understanding but not the target misconception, and have profile  $\alpha_j = [1, 0]$ .

Table 15.2 provides the three Q matrices based on models with two misconceptions. The Q matrices for models 2A, 2B, and 2C follow the same Q matrix pattern

**Table 15.1** Q Matrix specification for model with one misconception attribute

Response Type	Skill	Misconception
Incorrect	0	0
Misconception	0	1
Correct	1	0

**Table 15.2** Q Matrix specifications for models with two misconception attributes

Q Matrix	Response Type	Items 1–5			Items 6–12		
		S	M1	M2	S	M1	M2
Model 2A	Incorrect	0	0	N	0	N	0
	Misconception	0	1	N	0	N	1
	Correct	1	0	N	1	N	0
Model 2B	Incorrect	N	N	N	N	N	N
	Misconception	N	1	N	N	N	1
	Correct	1	N	N	1	N	N
Model 2C	Incorrect	0	N	N	0	N	N
	Misconception	N	1	N	N	N	1
	Correct	1	0	N	1	N	0

Note: S = desired understanding attribute, M1 = first misconception attribute, M2 = second misconception attribute.

as that in Table 15.1, but now include two separate misconception attributes. Misconception 1 relates to squares and rectangles and is measured by items 1–5 while Misconception 2 relates to parallelograms and is measured by items 6–12. The attribute vectors have the form  $\alpha_j = [I_s, I_{M1}, I_{M2}]$ , where the elements indicate possession (or not) of the skill, first misconception and second misconception, respectively. For these Q matrices, items 1–5 have one type of Q matrix entry and items 6–12 have a second type of Q matrix entry. In Table 15.2 the “S” column represents the desired skill, “M1” represents the first misconception and “M2” represents the second misconception.

Model 2A in Table 15.2, for example, extends the Q matrix shown in Table 15.1, and maintains the same interpretations. The first row for items 1–5, which is  $q_{ih} = [0, 0, N]$  indicates that selecting an incorrect option for one of the first five items is evidence that the student possesses neither the desired skill attribute nor the first misconception; this item does not provide evidence about the second misconception, indicated by the “N” entry in the third position. Interpretations for the Q matrix entries for items 6–12 are analogous to those for items 1–5, but relate to the second misconception rather than to the first misconception.

There are other ways the Q matrix could be constructed using the 0/1/N coding that are also consistent with the theory underlying test development. Models 2B and 2C in Table 15.2 illustrate two other possible ways to capture the hypothesized relationships between attributes and response options. Contrasting these different Q matrices highlights important issues that arise when interpreting the model that each Q matrix represents.

In a Q matrix for dichotomous items, the response probability for an item depends on only a single row. In other words, one need only look across the row of a Q matrix to understand how the attributes and the 0/1 entries will impact responding and make comparisons across possible Q matrices. In the GDCM-MC Q matrix, however, the response probabilities depend on the values in both the rows and the columns (within items). This is a result of the fact that the likelihood of an examinee

choosing each option depends not only on the attributes linked to that option, but also on the likelihood of an examinee selecting one of the other options. As a result, although Models 2A and 2C share the same Q matrix row for correct options,  $q_{ih} = [1, 0, N]$ , the probability of an examinee with a given attribute profile selecting the correct option under these models could differ.

To illustrate this dependence, consider the Q matrix for Model 2C for items 1–5. The row for incorrect options is  $[0, N, N]$ . This implies that a student who possesses the correct understanding is less likely to choose this option than a student who does not possess the understanding. The “N” in position two implies that possession of the first misconception should not matter for selecting this option. However, because the rows for the misconception and correct options contain “0” and “1” in position two, respectively, possession of the first misconception will impact the likelihood of selecting those options. When computing the probability that an examinee selects the incorrect option then, the computation will involve  $S_{i,\alpha_1}$ , which does depend on possession of the first misconception. In practical terms, this means that possessing the first misconception can indirectly influence the likelihood of selecting an incorrect option even though the Q matrix entry corresponding to the incorrect option and the first misconception is “N.”

To see this explicitly, consider the model-implied probabilities for item 1 using Model 2C and the two profile vectors  $\alpha_1 = [0, 1, 0]$  and  $\alpha_2 = [0, 0, 0]$ . To write the probability of a correct response in terms of Eq. 15.5, let  $h = 0$  represent an incorrect response. Omitting item subscripts on the  $\pi$  and  $r$  parameters for clarity, the model-implied probability of a correct response for  $\alpha_1$  is

$$P(X_1 = 0 | \alpha_1) = \frac{\pi_0}{\pi_0 + \pi_1 + \pi_2 r_{21} r_{22}} \quad (15.9)$$

if the sum in the denominator is greater than or equal to 1 and

$$P(X_1 = 0 | \alpha_1) = \pi_0 + (1 - [\pi_0 + \pi_1 + \pi_2 r_{21} r_{22}]) \left( \frac{1}{3} \right) \quad (15.10)$$

otherwise. For the profile  $\alpha_2$ , the probability is

$$P(X_1 = 0 | \alpha_2) = \frac{\pi_0}{\pi_0 + \pi_1 r_{12} + \pi_2 r_{21}} \quad (15.11)$$

if the sum in the denominator is greater than or equal to 1 and

$$P(X_1 = 0 | \alpha_2) = \pi_0 + (1 - [\pi_0 + \pi_1 r_{12} + \pi_2 r_{21}]) \left( \frac{1}{3} \right) \quad (15.12)$$

otherwise. Hence although the term  $\pi_0$  appears in both formulas, the predicted probabilities of selecting the incorrect response option will in general not be equal unless the following equality holds:

$$\pi_0 + \pi_1 + \pi_2 r_{21} r_{22} = \pi_0 + \pi_1 r_{12} + \pi_2 r_{21}. \quad (15.13)$$

Although this is possible, it is unlikely. This highlights the fact that even when a response option has an “N” entry for a particular attribute, possession of that attribute can still indirectly influence the probability of selecting the given option.

Interpretations of the cognitive processes operationalized by the Q matrices in Tables 15.1 and 15.2 must take into account both the rows and columns of the Q matrix for all options of an item. Doing so, we can propose three heuristic interpretations for the Q matrices represented in Tables 15.1 and 15.2. Model 2A implies a “strong” or “consistent” response process, in which examinees are unlikely to choose a response option inconsistent with their attribute profile. If an examinee selects a response inconsistent with their attribute profile, there is no clear hypothesis built into the model about how they would be likely to err. The same interpretation applies to the model in Table 15.1, except that it involves only a single misconception.

Model 2B implies a much less consistent model of responding because of the increased use of “N” terms. Those who possess the desired understanding (first attribute) are more likely to select the correct answer and those who possess the misconception measured by an item (second or third attribute, depending upon item) are more likely to select the misconception options. However, there is no a priori, built-in structure determining which of the other response options examinees might select if they do not select the option most consistent with their profile.

Model 2C implies more consistency of responding than Model 2B, but also implies that certain types of inconsistencies are more likely than others. For example, if those who possess the desired understanding (first attribute) do not select the correct response they are more likely to select a misconception rather than another incorrect response. This is because the 0 in the incorrect response row of the first column implies they are unlikely to select these options. Similarly, if someone with the first misconception does not select the misconception response (on items 1–5), they are more likely to select a miscellaneous incorrect response than a correct response, because the correct response contains a “0” for the second attribute while the incorrect response row contains an “N.” The model implies less about how those without either attribute will respond, although they will tend to respond with incorrect responses because they do not “mismatch” these option vectors, whereas they “mismatch” at least one element of the other two response option vectors.

In summary, the three models could be assigned conceptual labels such as consistent responding (Model 1, Model 2A), inconsistent responding (Model 2B) and predictable errors (Model 2C). Note that these Q matrices only use information about whether a response is coded as an incorrect, misconception, or correct response. Although this modeling approach contains substantially more information about student responses than using only information about whether responses are

correct or incorrect, it leaves out additional information about the exact response pattern on each item. This modeling approach was selected because it utilizes information in a way that is directly analogous to the scoring key proposed by the test developers.

## Data

The data used come from a national sample of 2,011 students collected as part of a randomized controlled trial designed to evaluate the efficacy of using the Diagnostic Geometry Assessment tests and associated curricular materials as formative classroom assessments (Masters, 2014). Students were selected from approximately 45 classrooms and are primarily in 6th–8th grade. The analysis uses student responses to the shape properties pre-test administered at the start of the study. Students who left entire items blank were removed from the dataset. Overall, 1927 students had complete responses to all questions, while 84 had partially complete responses to between 1 and 7 of the 12 items, and were included in the analyses. Partially complete responses occur when an item requires students to respond to multiple binary choices, but the student does not respond to all of them. Partially complete responses could never be scored as correct, but could be scored as representing a misconception or other incorrect response.

The average number of correct responses was 4.70 (min = 0, max = 12,  $SD = 3.43$ ) and the average number of “misconception” responses was 3.85 (min = 0, max = 12,  $SD = 2.85$ ). Table 15.3 provides classical item statistics for the items, including the percent of correct, misconception and incorrect responses selected as well as

**Table 15.3** Classical item statistics

Item	Correct Response			Misconception Response		
	p	Item-total	Item-rest	p	Item-total	Item-rest
1	0.19	0.56	0.47	0.43	0.42	0.26
2	0.39	0.62	0.52	0.26	0.47	0.34
3	0.25	0.58	0.48	0.55	0.46	0.31
4	0.30	0.66	0.58	0.33	0.53	0.39
5	0.21	0.52	0.43	0.49	0.44	0.28
6	0.69	0.62	0.52	0.27	0.64	0.54
7	0.51	0.69	0.60	0.25	0.62	0.51
8	0.38	0.63	0.54	0.36	0.58	0.45
9	0.32	0.55	0.45	0.22	0.44	0.31
10	0.33	0.58	0.48	0.15	0.58	0.49
11	0.57	0.66	0.57	0.28	0.55	0.43
12	0.57	0.70	0.61	0.27	0.60	0.49

Note: p is the proportion of students selecting the indicated response option type. Data are for  $n = 2,011$  test-takers.

item-total and item-rest correlations when scored for number correct and number misconception. The misconception scoring in Table 15.3 treats the items as measuring a single misconception attribute.

### ***Model Estimation and Evaluation***

Model estimation was carried out using a custom Markov Chain Monte Carlo (MCMC) algorithm written in FORTRAN, and described in greater detail by DiBello et al. (2015) and Hartz (2001). Chains of length 30,000 with burn-in periods of 24,000 iterations were used. Expected a posteriori (EAP) estimates, operationalized as the mean parameter estimate for the post burn-in iterations, were used for item parameters, examinee latent profile vectors (where examinee  $j$  was classified as possessing attribute  $k$  if  $\hat{P}[\alpha_{jk} = 1] > 0.5$ ), and estimated marginal proportions of examinees possessing each attribute.

The parameter estimates for each Q matrix were evaluated using three sets of information. First, convergence was assessed using post burn-in chain plots for each parameter. Second, global model fit was evaluated using Akaike information criterion (AIC) and Bayesian information criterion (BIC) indices, while local fit was evaluated using the option-fit statistic,  $D_{ih}$ , described in DiBello et al. (2015, p. 72).<sup>3</sup> The option fit index characterizes the discrepancy between the observed frequency of each response option with the model-predicted frequency, taking into account the model-predicted examinee classifications and estimated parameters. Third, using a similar procedure to that described in DiBello et al., the mean absolute difference between model parameter estimates based on real and simulated dataset were compared.<sup>4</sup> The simulated data were also used to estimate correct classification rates for each attribute.

Overall model interpretability and utility were evaluated using model-implied response option probability plots for each latent profile and item discrimination indices using the  $d_{ik}$  statistic described in DiBello et al. (2015) and an alternative index,  $d'_{ik}$ , described below. The discrimination index  $d_{ik}$  is calculated for each item  $i$  by attribute  $k$  combination as:

<sup>3</sup>The expression for AIC is  $AIC = -2LL + 2P$ , where  $LL$  is the log likelihood of the model and  $P$  is the number of item parameters; the expression for BIC is  $BIC = -2LL + \log(n)P$  where  $n$  is the sample size (Agresti, 2013). Here the  $LL$  value based on the final parameter estimates (posterior means) and predicted examinee classifications were used in the calculation of the formulas and  $P$  is based on the number of item parameters estimated.

<sup>4</sup>Briefly, a set of  $N = 2,011$  (the original sample size) examinees and associated responses to the 12 test items were simulated, treating the estimated item parameters and examinee classifications as true population values. The GDCM-MC model parameters were re-estimated based on the simulated responses and compared to the original GDCM-MC estimates to evaluate stability of the parameter estimates and correct classification rates. Estimates based on this simulation procedure are likely to provide an upper bound to the parameter stability and classification accuracy rates, because they assume that the model is correctly specified (the original model is used to generate the simulated data).

$$d_{ik} = \left( \frac{1}{2} \right) \max_{\beta_{-k}} \left\{ \sum_{h=1}^{H_i} \left| \hat{P}(X_i = h | \alpha_k = 1, \beta_{-k}) - \hat{P}(X_i = h | \alpha_k = 0, \beta_{-k}) \right| \right\}, \quad (15.14)$$

where  $\beta_{-k}$  is a vector of attribute indicators with the  $k^{th}$  attribute removed and  $P(X_i = h)$  is the probability of selecting option  $h$  on item  $i$ . The differences

$$\hat{P}(X_i = h_i | \alpha_k = 1, \beta_{-k}) - \hat{P}(X_i = h_i | \alpha_k = 0, \beta_{-k}) \quad (15.15)$$

compare the model-predicted probability of selecting option  $h_i$  for two examinees who differ only in whether they possess attribute  $k$ . The  $d_{ik}$  index takes on values from 0 to 1, with 1 indicating greater discrimination. Here a new discrimination index,  $d'_{ik}$ , is defined for item  $i$  and facet  $k$  as:

$$d'_{ik} = \max_h \left\{ \frac{\left[ \sum_{\beta_{-k}} \left[ \hat{P}(X_i = h | \alpha_k = 1, \beta_{-k}) - \hat{P}(X_i = h | \alpha_k = 0, \beta_{-k}) \right] \right]}{2^{K-1}} \right\}. \quad (15.16)$$

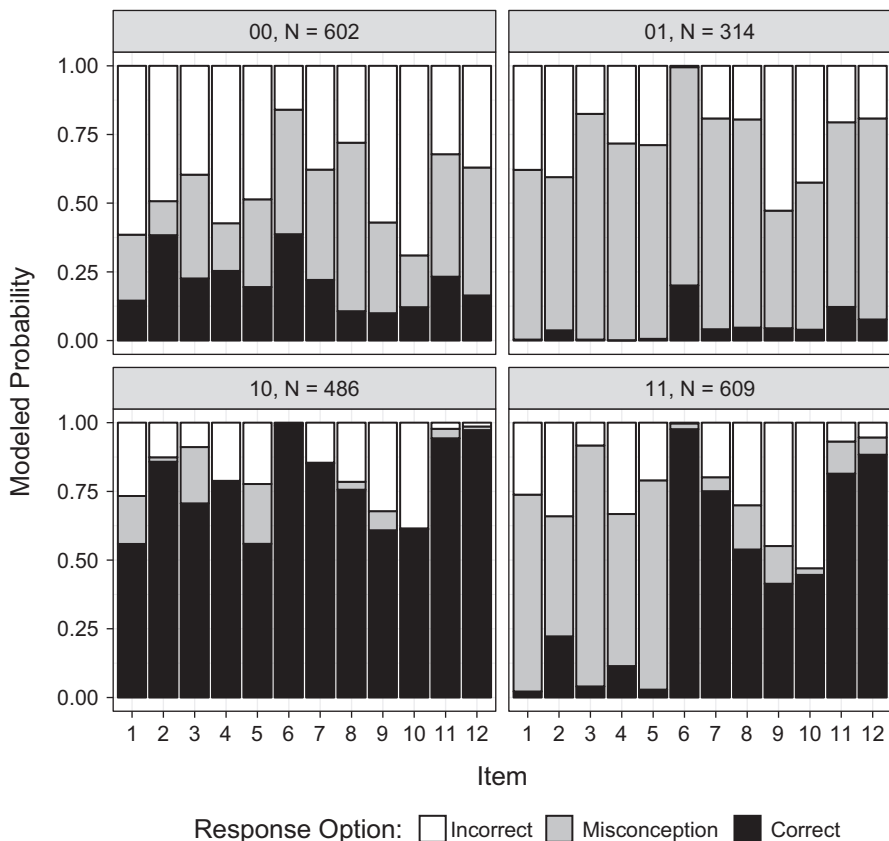
Because there are  $2^{K-1}$  possible profiles when excluding attribute  $k$ , this index also takes on values between 0 and 1, with values near 1 indicating greater discrimination. Although similar, these indices are sensitive to slightly different types of items. The  $d_{ik}$  index takes on values near 1 when at least one attribute profile has consistently large differences in expected response probabilities across the options of an item. In contrast, the  $d'_{ik}$  index takes on values near 1 when at least one item option has a large difference in expected response probabilities for examinees who differ only by attribute  $k$ , averaging across all possible profiles. Both indices can be used to provide information about how much information different items provide about each attribute.

## Results

Examination of item parameter chain plots and global fit statistics indicated a lack of fit for the single misconception model (Model 1), but the results for this model are presented as further evidence suggesting the need for a second misconception attribute. Figure 15.2 shows the model-predicted probabilities of selecting each response type for each item for Model 1, separately for all four possible attribute profiles. Within each panel, the stacked bar represents a single item; the white portion of the bar represents the probability of selecting an incorrect response option, the gray portion represents the probability of selecting a misconception response option, and the black portion represents the probability of selecting a correct response option.

In the top left panel of Fig. 15.2, for example, the item probabilities are shown for the profile  $\alpha_j = [0, 0]$  indicating a test-taker with neither the desired understanding





**Fig. 15.2** Item probability plots by profile for Model 1 (Note: Panel headings indicate profile and number of test-takers classified into the profile. Profiles are represented as 2-dimensional vectors of binary indicators, where the first element indicates possession of the skill attribute and the second element indicates possession of the misconception.)

nor the target misconception; there were 602 examinees (approximately 30%) classified into this profile. The predicted probabilities of selecting incorrect, misconception, or correct responses to item 1 for these examinees were 0.62, 0.24, and 0.15, respectively. The top right panel (profile of  $\alpha_j = [0, 1]$ ) represents examinees classified as having the single overall misconception, and who have a high likelihood of selecting misconception options for all items. The bottom left panel shows those with the desired understanding and not the target misconception, and indicates high probabilities of selecting the correct response options. The lower right panel, however, shows a clear separation between items 1–5 and items 6–12. This panel is for the profile  $\alpha_j = [1, 1]$ , which represents examinees showing evidence of both the desired skill and target misconception. This panel appears to suggest that there are a group of examinees that respond differently to items 1–5 than to items 6–12 in terms

**Table 15.4** Global model fit statistics

Model	LL	AIC	BIC	Parm.	MAD			Opt. Fit
					$\pi$	$r$	$p_k$	
Model 1	-17781.71	35779.43	36384.92	108	0.04	0.05	0.01	0.02
Model 2A	-16399.86	33015.71	33621.20	108	0.05	0.06	0.01	0.03
Model 2B	-17485.50	35091.00	35427.38	60	0.04	0.02	0.01	0.06
Model 2C	-16387.95	32943.89	33414.83	84	0.04	0.03	0.01	0.03

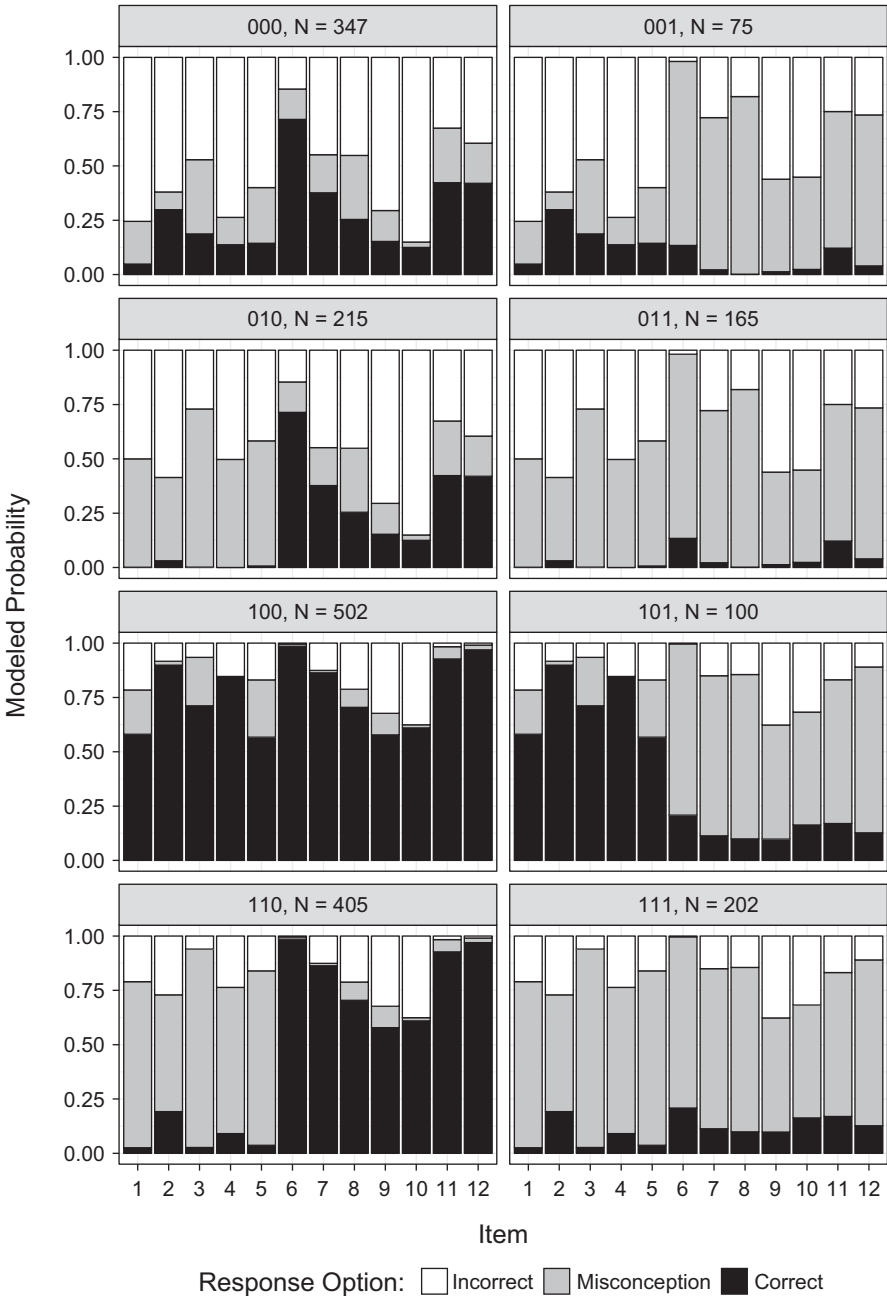
Note: *LL* = log likelihood, *AIC* = Aikake Information Criterion, *BIC* = Bayesian Information Criterion, *Parm* = number of parameters, *Opt. Fit* = average option fit statistic, *MAD* = mean absolute difference.

of their likelihood of selecting correct or misconception responses, and highlights the need for two misconception dimensions.

Table 15.4 summarizes overall model fit for each of the four models. Item parameter estimate chain plots for Models 2A, 2B, and 2C suggested that some of the parameters in Model 2A were not well estimated and possibly unnecessary. Parameter convergence for Model 2B was more consistent, but the model fit statistics in Table 15.4 suggest this model does not fit the data as well as Models 2A and 2C. Table 15.4 indicates that Model 2C had the lowest AIC and BIC values (lower values indicate better model fit). The mean absolute difference (MAD) between the model parameter estimates and the estimates based on the simulated data were relatively similar across models, although they tended to be lowest for Model 2B. The average option fit values indicate an average deviation between fitted and observed values, with lower values indicating better fit; the option fit statistic thus suggests Model 1 has the best fit, followed by Models 2A and 2C (which are similar) and then significantly worse fit for Model 2B. Taken together, Model 2C appeared to represent the best balance of overall model fit and sufficient information for stable parameter estimates to converge.

Figure 15.3 shows the model-predicted probabilities of selecting each response type for each item for Model 2C, separately for all  $2^3 = 8$  possible attribute profiles. The plot follows the same structure as Fig. 15.2. The heading of each panel indicates the attribute profile and the number of examinees (out of 2,011) classified into each profile; the  $p_k$  values, the estimated marginal proportion of students with each attribute, were 0.60, 0.49 and 0.28 for the desired understanding, first misconception, and second misconception, respectively. Again, the visual nature of the plots helps to provide an understanding of the types of response patterns that each profile represents, and hence of the types of response patterns found in the data.

In Fig. 15.3, the second panel from bottom on the left, for example, shows responses for the  $\alpha_j = [1, 0, 0]$  profile, which represents possession of the desired skill and neither misconception. The modeled probabilities indicate systematically high likelihoods of selecting the correct response options, and the model classified 502 students (approximately 25%) as responding with this profile. The second panel



**Fig. 15.3** Item probability plots by profile for Model 2C (Note: Panel headings indicate profile and number of test-takers classified into the profile. Profiles are represented as 3-dimensional vectors of binary indicators, where the first element indicates possession of the skill attribute, the second element indicates possession of the first misconception and the third element indicates possession of the second misconception.)

**Table 15.5** Item discrimination and guessing values

Item	Type	$d$			$d'$			Guessing
		S	M1	M2	S	M1	M2	
1	5 binary	0.54	0.56		0.42	0.43		0.00
2	4 MC	0.60	0.71		0.43	0.49		0.21
3	5 binary	0.52	0.69		0.31	0.54		0.09
4	4 binary	0.71	0.75		0.43	0.52		0.14
5	3 binary	0.43	0.54		0.34	0.43		0.00
6	4 MC	0.27		0.78	0.17		0.74	0.10
7	4 MC	0.49		0.75	0.29		0.63	0.14
8	8 binary	0.45		0.67	0.27		0.60	0.20
9	5 binary	0.43		0.48	0.28		0.36	0.03
10	6 binary	0.49		0.51	0.35		0.45	0.01
11	4 MC	0.50		0.76	0.28		0.53	0.35
12	4 MC	0.55		0.84	0.32		0.63	0.22

Note: *MC* = multiple-choice,  $d$  = discrimination index,  $d'$  = alternate discrimination index,  $S$  = desired understanding attribute,  $M1$  = first misconception attribute,  $M2$  = second misconception attribute.

from top on the right shows the profile  $\alpha_j = [0, 1, 1]$ , possessing both misconceptions but not the desired skill. Here we see generally high likelihoods of selecting misconception responses, and higher probabilities of selecting incorrect response options for some items (particularly items 2, 9, and 10). For other profiles, such as  $\alpha_j = [1, 0, 1]$ , examinees tended to answer items on the first half of the test correctly, but tended to select misconception responses on the second half of the test. This highlights the importance of considering an examinee’s overall profile classification, not only their classification on individual attributes.

The bottom right panel of Fig. 15.3 represents the profile  $\alpha_j = [1, 1, 1]$ . Nominally, this represents a student who possesses both the desired skill and both misconceptions. The probability plot shows that these students consistently select misconception responses, but tend to select other incorrect responses much less often than students in the  $\alpha_j = [0, 1, 1]$  profile. Hence, although the attribute profile is not necessarily one that would be theoretically anticipated, there is a clear interpretation for the type of response pattern represented by this profile.

The classification of students into these attribute profiles may be useful instructionally for teachers seeking to better understand their students’ reasoning. It is important, however, to ensure that such classifications are reliable. The simulation-based approach described above was used to estimate the classification accuracy for each of the individual attributes and for the overall profile classifications. The simulation-based accuracy rates were high: 0.93, 0.93, and 0.98 for the skill, first misconception, and second misconception attributes respectively, and 0.85 for the overall classification into one of the eight profiles. Assuming uniform prior probabilities for each attribute, then  $\kappa_n$ , a chance-corrected overall classification accuracy statistic similar to Cohen’s  $\kappa$  (Brennan & Prediger, 1981) is approximately 0.83.

**Table 15.6** Original and optimal cutscores with agreement rates

Attribute	Items	Original Cutscore	Optimal Cutscore	Agreement	GDCM N	Optimal N
Skill	12	9	8	0.95	502	435
Misconception 1	5	2	2	0.89	987	1184
Misconception 2	7	3	3	0.95	542	614

Note: GDCM N = number of test-takers classified as possessing attribute by GDCM, Optimal N = number of test-takers classified as possessing attribute using sum scores with Optimal Cut.

This represents a high level of classification accuracy, particularly for a test used for formative assessment purposes, and was higher than the classification accuracy rates using the sum scores classifications.<sup>5</sup>

Table 15.5 presents the  $d_{ik}$  and  $d'_{ik}$  discrimination values, item type and average model-predicted probability of guessing for each item. In Model 2C items 1–5 have a discrimination value of 0 for misconception two and items 6–12 have a value of 0 for misconception one; these cells are thus left blank. The two discrimination indices are not directly comparable in magnitude, but are highly correlated, suggesting they tend to provide similar results. Both indices suggest item 6 provides the least information about the skill attribute, but they present different pictures about the level of information provided by items 1, 2, and 4. While the  $d'_{ik}$  index suggests these items provide similar amounts of information regarding the skill attribute, the  $d_{ik}$  index suggests that item 4 provides substantially more information than items 1 and 2. Both indices suggest the items provide more information about the misconception attributes than about the skill attributes. The average guessing probabilities were higher on average for the multiple-choice items ( $M = 0.20$ ,  $SD = 0.10$ ) than for the binary choice items ( $M = 0.07$ ,  $SD = 0.08$ ). When writing additional items, these results suggest that items similar to item 6 may not provide very helpful information, while items similar to 2, 4, and 12 would be more promising to emulate.

The GDCM-MC results were also used to evaluate the classification cutscores suggested by the test developers. For each of the three attributes, the observed score cutscore that would maximize agreement with the GDCM-MC model-based classifications was identified. Table 15.6 displays the original observed-score cutscores (“Original Cutscore”) used to classify test takers as possessing either the skill and/or the misconceptions measured by items 1–5 and 6–12. These cutscores were based on the 35 and 75% rules described above. The “Optimal Cut” column reports the sum-score cutscores that would result in the highest agreement with the GDCM-MC classifications. The classification agreement rate using these optimal cutscores and the GDCM-MC classifications are reported under the column “Agreement.”

<sup>5</sup>Comparing the sum-score classifications of simulated data to the profile classifications in the original data based on the GDCM-MC model, agreement rates are 0.58, 0.86, 0.97, and 0.46 for the first attribute, second attribute, third attribute and overall profile, respectively. Comparing the sum-score classifications of the simulated data to the sum-score profile classifications in the original data, the agreement rates are 0.89, 0.79, 0.92, and 0.66, respectively. The GDCM-MC classification accuracy rates are higher in both instances.

Agreement rates were very high; approximately 0.95 for the skill and second misconception and 0.89 for the first misconception (based on only 5 items). These analyses support the original cutscores identified earlier with one small exception. The number of correct responses required to be classified as a “knower” is higher (9 out of 12) in the original guidelines than the optimal cutscore identified here (8 out of 12). Depending upon whether false positives or false negatives are of greater concern, these cutscores could be modified. Additional calculations could be used to select cutscores minimizing one or the other type of error.

## Discussion

There may be some uncertainty about how best to categorize the validity evidence presented in this chapter using the categories outlined in the *Standards*. On the one hand, the results are based primarily on modeling the observed item response data and could fall under the internal structure heading. On the other hand, the match between the GDCM-MC framework and test design process generated results that can inform our understanding of the response processes examinees may be using on the shape properties test. Although the analyses do not provide direct evidence about examinees’ response processes, they do allow researchers to evaluate whether proposed inferences regarding response processes are consistent with the model results. From that perspective, these results provide additional support for the proposed interpretations and uses of the shape properties test. The results also suggest directions for future validity studies after identifying response patterns in the data consistent with scores measuring two dimensions of the targeted misconception rather than one. The GDCM-MC model-based classifications also supported the original sum-score cutscores (Masters, 2010; Russell et al., 2009), with only a slight modification to the cutscore used to identify students reasoning correctly. There are several factors that contributed to the success of these analyses, and that highlight some of the strengths and limitations of using the GDCM-MC (or a similar) framework to validate distractor-driven tests.

First, the design and intended use of the shape properties test fit naturally with the GDCM-MC modeling framework. The importance of this fact should not be overlooked. DCMs are often “retrofitted” (Gierl & Cui, 2008) to tests that are not designed to yield diagnostic information about multidimensional attributes, and can produce unsatisfying results. While it is preferable to develop diagnostic assessments using a process that includes interaction between test design, administration and psychometric modeling (Roussos et al., 2010), the analyses here suggest distractor-driven tests may serve as one area where retrofitting DCMs could be productive. By analyzing existing distractor-driven tests, researchers could readily compare results based on recently developed DCMs for nominal responses (e.g., Bradshaw & Templin, 2014; de la Torre, 2009a) or less parametric techniques such as cluster analysis (e.g., Chiu, Douglas, & Li, 2009; Chiu & Köhn, 2015) without first needing to develop new assessments.

Other strengths of the GDCM-MC framework include model flexibility, the reliance on readily available item response data, and the rich array of model-based results. The ability to model multiple attributes with complex links to each response option greatly expands the types of cognitive models that can be fit and compared. But these strengths do come at the cost of a substantially more complex Q matrix construction and model estimation processes than those used for other psychometric models, including DCMs for binary items. Given the centrality of Q matrix specification to DCMs, this is an important concern for applied use. Furthermore, users should be cautious about the complexities involved in the necessary and sufficient conditions for model identification for the GDCM-MC (DiBello et al. 2015).

From a validity perspective, more direct evidence regarding examinee response processes and GDCM-MC classifications would certainly be useful. Verbal reports that elicit students' reasoning as they complete the test items, possibly using protocol analysis (e.g., Ericsson & Simon, 1993), could be used to verify whether students appear to be reasoning in the ways predicted by the GDCM-MC classifications. Leighton and Gierl (2007b) provide an extended discussion on the use of verbal reports to develop and validate diagnostic assessments. Alternatively, the GDCM-MC classifications could be triangulated by making classifications based on a different method, such as student interviews, or a separate test, and then evaluating the level of agreement with the GDCM-MC classifications. Similar studies were carried out on the shape properties test using the earlier unidimensional misconception scoring (Masters, 2012a, 2012b). These studies found that while trained researchers' judgments about student misconceptions during interviews had moderate agreement with the shape properties test scores, teachers' judgments about student misconceptions tended to have low agreement with the test scores. It would be useful to determine whether agreement rates are higher for classifications that are based on GDCM-MC scores and a two-dimensional misconception space. The potential for this type of iterative investigation, between qualitative analyses of student responding and large-scale psychometric modeling, further underscores the value of the GDCM-MC framework in this context.

Of course, a primary goal of the GDCM-MC framework is to improve assessment practices and learning outcomes, either by improving our understanding of test score meaning or generating more useful test scores. Masters (2014) provides preliminary evidence that using the Diagnostic Geometry Assessment tests and materials can improve student outcomes. Future studies could evaluate whether including the new GDCM-MC scores and associated information further improve these outcomes or provide teachers with additional instructionally useful information. Similarly, it would be useful to know which of the GDCM-MC results are most useful to test developers and item writers who construct distractor-driven tests. Hopefully, the pursuit of such studies will improve the utility both of diagnostic assessments and of the GDCM-MC and other related psychometric models.

**Acknowledgement** The authors gratefully acknowledge the feedback and software provided by William Stout, Louis DiBello and Robert Henson for this work. The Diagnostic Geometry Assessment Project data was generously shared by Jessica Masters and was collected with funding from an Institute of Education Sciences Grant (#R305A080231).



## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5, 7–74. <https://doi.org/10.1080/0969595980050102>
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, 79, 403–425. <https://doi.org/10.1007/s11336-013-9350-4>
- Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41, 687–699. <https://doi.org/10.1177/001316448104100307>
- Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, 11, 33–63. [https://doi.org/10.1207/s15326977ea1101\\_2](https://doi.org/10.1207/s15326977ea1101_2)
- Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74, 633–665. <https://doi.org/10.1007/s11336-009-9125-0>
- Chiu, C.-Y., & Köhn, H.-F. (2015). Consistency of cluster analysis for cognitive diagnosis: The DINO model and the DINA model revisited. *Applied Psychological Measurement*, 39, 465–479. <https://doi.org/10.1177/0146621615577087>
- de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, 33, 163–183. <https://doi.org/10.1177/0146621608320523>
- de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115–130. <https://doi.org/10.3102/1076998607309474>
- DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement*, 39, 62–79. <https://doi.org/10.1177/0146621614561315>
- DiBello, L. V., Henson, R. A., Stout, W. F., & Roussos, L. A. (2014). *Under the hood: Applying the GDCM-MC family of diagnostic models for multiple choice option-based scoring to investigating the Diagnostic Geometry Assessment*. Presented at the Annual Meeting of the American Educational Research Association, Philadelphia.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press.
- Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research & Perspective*, 6, 263–268. <https://doi.org/10.1080/15366360802497762>
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 301–321. <https://doi.org/10.1111/j.1745-3984.1989.tb00336.x>
- Hartz, S. M. (2001). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality* (Doctoral dissertation). Urbana, IL: University of Illinois at Urbana-Champaign.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141. <https://doi.org/10.1119/1.2343497>
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258–272. <https://doi.org/10.1177/01466210122032064>

- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. <https://doi.org/10.1111/jedm.12000>
- Leighton, J. P., & Gierl, M. J. (2007a). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.
- Leighton, J. P., & Gierl, M. J. (2007b). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 146–172). New York, NY: Cambridge University Press.
- Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions and applications*. Charlotte, NC: Information Age Publishing Inc.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448. <https://doi.org/10.3102/0013189X07311286>
- Luecht, R. M. (2007). Using information from multiple-choice distractors to enhance cognitive-diagnostic score reporting. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 319–340). New York, NY: Cambridge University Press.
- Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75*, 491–511. <https://doi.org/10.1177/0013164414539162>
- Masters, J. (2010). *Diagnostic geometry assessment project technical report: Item characteristics*. Chestnut Hill, MA: Lynch School of Education Boston College.
- Masters, J. (2012a). *Diagnostic Geometry Assessment project: Validity evidence* (Technical Report). Measured Progress Innovation Lab
- Masters, J. (2012b). *The validity of concurrently measuring students' knowledge and misconception related to shape properties*. Presented at the annual meeting of the American Educational Research Association, Vancouver, BC.
- Masters, J. (2014). *The diagnostic geometry assessment system: Results from a randomized controlled trial*. Presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.
- Minstrell, J. (2000). Student thinking and related assessment: Creating a facet assessment-based learning environment. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP*. Washington, DC: National Academy Press.
- Roussos, L. A., DiBello, L. V., Henson, R. A., Jang, E., & Templin, J. (2010). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 35–69). Washington, DC: American Psychological Association.
- Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007a). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). New York, NY: Cambridge University Press.
- Roussos, L. A., Templin, J. L., & Henson, R. A. (2007b). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*, 293–311. <https://doi.org/10.1111/j.1745-3984.2007.00040.x>
- Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96. <https://doi.org/10.1177/0013164407301545>
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Russell, M., O'Dwyer, L. M., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods, 41*, 414–424. <https://doi.org/10.3758/BRM.41.2.414>

- Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, 35, 265–296. [https://doi.org/10.1002/\(SICI\)1098-2736\(199803\)35:3<265::AID-TEA3>3.0.CO;2-P](https://doi.org/10.1002/(SICI)1098-2736(199803)35:3<265::AID-TEA3>3.0.CO;2-P)
- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing Inc..
- Smith, J. P., III, diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, 3, 115–163. [https://doi.org/10.1207/s15327809jls0302\\_1](https://doi.org/10.1207/s15327809jls0302_1)
- Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345–354. <https://doi.org/10.1111/j.1745-3984.1983.tb00212.x>
- Templin, J. (2016). Diagnostic assessment: Methods for the reliable measurement of multidimensional abilities. In F. Drasgow (Ed.), *Technology and testing* (pp. 285–304). New York, NY: Taylor & Francis.
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, 30, 251–275. <https://doi.org/10.1007/s00357-013-9129-4>
- Vinner, S., & Hershkowitz, R. (1980). Concept images and common cognitive paths in the development of some simple geometrical concepts. In R. Karplis (Ed.), *Proceedings of the Fourth International Conference for the Psychology of Mathematics Education* (pp. 177–184). Berkeley, CA.
- Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26, pp. 45–79). Amsterdam, The Netherlands: Elsevier Science B.V.
- Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences*. New York, NY: Springer International Publishing. Retrieved from <http://link.springer.com/10.1007/978-3-319-07794-9>

# Chapter 16

## Understanding Test-Taking Strategies for a Reading Comprehension Test via Latent Variable Regression with Pratt's Importance Measures

Amery D. Wu and Bruno D. Zumbo

### Introduction

#### *Test-Taking Strategy and Validity*

Scholars and practitioners in language assessment often call for research and validation practices that look into the processes of test-taking. Anderson, Bachman, Perkins, and Cohen (1991), pointed out the importance of gathering information on test-taking processes as part of construct validation. A process-based approach focuses on test-takers' active engagement with test questions. This advocacy for a process-based approach to validation emerged from dissatisfaction with the traditional outcome-based approaches to validation in language testing. The outcome-based approaches centered on the product of a test (i.e., scores) and a psychometric method that studied the product's correlations with other outcome measures as evidence of validity.

The outcome-correlation approach to validation is limited in revealing the processes that the test-takers may employ in a test (Allen, 2003; Cohen & Upton, 2006; Farr, Pritchard, & Smitten, 1990; Newton, 2016; Roizen, 1984). Outcome-correlation approaches to validation investigate test-takers' scores rather than their reaction and response to the test stimuli (e.g., the strategies and skills they employ in the test), which often is the ability a test actually intends to measure (or avoid). Outcome-correlation approaches might neglect, say, test-wiseness strategies (e.g., choosing the longest option in a multiple-choice item) that might have contributed to a test-taker's score even though the contribution is through an unwanted skill.

---

A.D. Wu (✉) • 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: [amery.wu@ubc.ca](mailto:amery.wu@ubc.ca); [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca)

Historically, outcome-correlation approaches for validity evidence are tied to traditional conceptions of validity that are practiced through simple correlations to show construct validity (convergent and discriminant) and/or criterion validity (concurrent or predictive). Kane (2001) and Zumbo (2007) attest that evidence of correlations between test scores with other variables only provides a weak form of validity evidence. A stronger form of validity should provide an *explanation* for the test scores (Zumbo, 2005). In contrast to outcome-based correlational approaches to provision of validity evidence, process-based approaches to validity investigate how a test-taker engages with and responds to test items (Newton, 2016). This approach provides an explanation for test scores variation by process variables rather than simply a concurrence in test score variation between two variables.

This emphasis on validity evidence by tracking item responding processes was supported by validity theorist and psychometrician Samuel Messick (1995). He summarized six aspects of construct validity. One aspect involved investigating whether test responses provided empirical evidence to support a claim that relevant cognition and skills are actually engaged by the respondents. Correspondingly, *the Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) recognized the evidence based on understanding responding processes as one of the four major sources of validity evidence. “Theoretical and empirical analysis 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 the test-takers” (AERA, APA, & NCME, 2014, p. 15).

In this chapter, we study how test-takers’ self-reports of three types of test-taking strategy are associated with their performance in an English reading comprehension test comprised of text passages and multiple-choice items. In doing so, we characterize test-taking strategies in a reading test as processes and, in particular, agentive processes, that steer a test-taker’s engagement in taking the test. In our view, studying these agentive processes as an empirical method for score validation addresses the above mentioned calls for process-based approaches to provision of validity evidence.

This chapter is organized in four parts. First, we will provide a brief review on the typology of test-taking strategies. Second, we will describe our explanatory approach to data analysis for investigating test score variation through a latent variable regression assisted by Pratt’s importance measures. Third, through our analytical plan and an example data set from a reading comprehension pilot test, we provide a demonstrative study to illustrate how process variables of test-taking strategies explain test score variation. Finally, we will discuss the findings in terms of advantages of our analytical approach as well as suggestions for test design and validation.

## Literature on Types of Test-Taking Strategy

Since the 1980s, researchers in language testing have looked into the processes of test-taking using qualitative approaches (e.g., Abbott, 2006; Cohen, 1984; Cohen & Upton, 2006, 2007; Farr et al., 1990; Lee, 2011; Nevo, 1989; Roizen, 1984; Rupp,

Ferne, & Choi, 2006). Most of the time, a method of verbal report such as think-aloud cognitive interviews were used to collect information about test-takers' strategies. Cohen (2006) reviewed 25 years of research in this area, and concluded that there has yet to emerge a more unified theory for test-taking strategies. Such work remains a challenge. However, these studies have identified a broad range of strategies involved in taking reading comprehension tests. Historically, these strategies are classified to three major categories: (a) reading for comprehending meaning, (b) test-management, and (c) test-wiseness (see reviews in Cohen, 2006, 2012a, 2012b; Cohen & Upton, 2006, 2007; Lee, 2011).

These qualitative studies report that when engaging in reading for *comprehending meaning* (CM), test-takers are drawing from their repertoire of linguistic and cognitive skills while processing the information in the texts such as identifying indicators of key ideas and looking for markers of meaning (e.g., definitions and examples). Test-management strategies are deployed to engage with the structure of the test itself, such as selecting an option in a multiple-choice question or accounting for test time when attending to texts and items. One specific sub-area of test-management is score-maximizing strategy, a strategy often employed in a testing setting when the test results have consequential implications on the test-takers. When employing *score-maximization* (SM) skills, test-takers are driven to maximize test scores relying on any partial or fragmented understanding from processing the texts to answers questions (e.g., utilizing partial understanding to eliminate seemingly implausible multiple-choice options). Finally, when exercising *test-wiseness* (TW), test-takers will be trying to select a correct answer without necessarily engaging in any of the expected linguistic and cognitive processes.

Despite there being a generally accepted typology of test-taking strategies and occasional discussion of their implications of test-taking strategies on validity (Anderson, Bachman, Perkins, & Cohen, 1991; Cohen, 2006, 2012a; Farr et al., 1990; Roizen, 1984), investigating test-taking strategies as a validation tool for providing empirical evidence using quantitative methods is still rare in language assessment research. An exception was Purpura (1998) who, through the application of structural equation modeling (SEM) techniques, found differences existed in the ways that high and low ability test-takers used strategies in response to different test tasks, and that these different patterns of employment in strategy use had a significant impact on second language test performance. As we will illustrate in this chapter, it is fruitful to take advantage of this typology and extend it to score validation by quantitative methods.

## **An Explanatory Approach to Validation by Understanding Test-Taking Strategies**

In a recent study, Wu and Stone (2015) investigated strategies in taking an English reading comprehension test via a cognitive survey. The survey asked the participants to report their use of ten test-taking strategies that were frequently identified

by qualitative literature to tap the three types of strategies. Through both exploratory and confirmatory factor analyses, the findings supported the three-dimensional structure. In this chapter, we continue to investigate individuals' employment of these three types of test-taking strategies via the method of latent variable regression in the SEM framework. We further explain (account for) the score variation through the method of Pratt's importance measures.

In a latent variable regression, one posits a model stating that a set of observed explanatory variables (types of test-taking strategy) affects the outcome variable, a latent performance variable that is indicated by the observed item scores. In current practices, explanation of test score variation is usually carried out by a regression method for an observed outcome variable that is a composite score of a set of item scores (i.e., number correct sum score). The problem with the observed variable regression is that the estimated regression results are biased in the observed composite score. Although commonly known among statisticians, this estimation bias is unfortunately ignored in day-to-day practices of validation research (Zumbo, 2007). Latent variable regression is a more optimal method for understanding test score variation than an observed score regression because it takes into account the measurement errors in the observed test scores.

Once the latent variable regression is conducted, the method of Pratt's importance measures will be applied. The method was originally proposed for ordering the importance of the explanatory variables in multiple regression analyses (Pratt, 1987; Thomas, Hughes, & Zumbo, 1998). In this chapter, the method will be applied to the results of a latent variable regression from a SEM (Zumbo, 2007). The goal is to look into the importance of test-takers' reports of using the three above-mentioned types of strategy in relation to the latent test performance. Before we do that, we will first describe the SEM specification for the latent variable regression and the method of Pratt's importance measures.

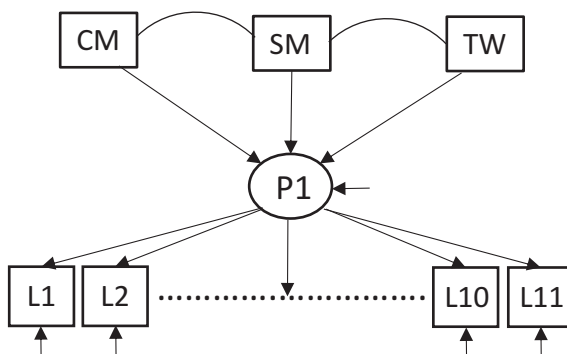
### *Latent Variable Regression*

Figure 16.1 depicts the latent variable regression using the SEM convention of a path diagram. The outcome variable (P1 depicted as an oval) is a latent variable for test performance on the task of reading a letter correspondence (see Table 16.1), which has 11 observed test items (L1–L11, each in a rectangle) as manifest variable indicators. The three inter-correlated explanatory variables are the three types of test-taking strategy of CM, SM, and TW. Each type of strategy is in a rectangle to show they are observed variables (the model specification will be described in more detail later in the analysis section). Then, the variation in the latent score P1 accounted for by the regression model (i.e., R-square) is partitioned by the three explanatory variables using the method of Pratt's measures (described next) to attribute the importance of each type of strategy.



**Fig. 16.1** SEM path diagram of latent variable regression of task performance on the three observed types of test-taking strategy

- L1 –L11: observed test items for the Letter task
- CM, SM, & TW observed test-taking strategy types
- P1: Latent performance for the Letter task



**Table 16.1** Summary the CELPIP-G reading pilot test and descriptive statistics of the sample data

Task type	Time (min)	# of Items	$\alpha$	M	SD
Task-1 Reading letter correspondences	11	11	0.74	0.76	0.22
Task-2 Reading to understand & apply diagram information	9	8	0.60	0.66	0.24
Task-3 Reading for factual information	10	9	0.81	0.59	0.30
Task-4 Reading for viewpoints	13	10	0.75	0.42	0.26
Whole test	43	38	0.90	0.61	0.21

Note. All descriptive statistics were calculated based on proportion correct score (number of correct answers divided by the corresponding number of questions);  $\alpha$  denotes Cronbach’s reliability estimate of internal consistency

### *Pratt’s Importance Measure*

It is sometimes recommended in regression textbooks and often followed in practice that the importance of a number of  $p$  explanatory variables be order by the absolute value of  $\beta_p$ , the standardized partial regression coefficient for the  $p$ th variable, because  $\beta_p$  is a standardized measure that circumvents the issues of incomparability. Incomparability is due to unstandardized regression coefficients are estimated for explanatory variables that have different units of measurement. This suggestion is problematic because it ignores the fact that the partial regression coefficient, whether it be standardized or not, is a measure of relationship between a specific explanatory variable with the outcome variable controlling for the relationships between the other  $(p - 1)$  explanatory variables with the outcome variable. For different explanatory variables, the set of  $(p - 1)$  controlled relationships will be different, and

hence their importance is not directly comparable. This problem is resolved by Pratt's (1987) approach.

Consider a linear multiple regression with one outcome variable of the form

$$Y = \hat{\beta}_{1X_1} X_1 + \hat{\beta}_{2X_2} X_2 + \dots + \hat{\beta}_{pX_p} X_p + U. \quad (16.1)$$

Pratt (1987) used an axiomatic approach to show that the unique importance of the  $p$ th explanatory variable could be expressed as  $\hat{\beta}_p \hat{r}_p$ , the product of its standardized regression coefficient and its simple correlation with the outcome variable  $Y$ . The  $R^2$ -standardized Pratt's measure,  $d_p$ , is later given by (Thomas et al. 1998) as,

$$d_p = \frac{\hat{\beta}_p \hat{r}_p}{R^2}. \quad (16.2)$$

Based on the fact that  $\sum_{p=1}^w \hat{\beta}_p \hat{r}_p = R^2$ , it follows that  $\sum_{p=1}^w \frac{\hat{\beta}_p \hat{r}_p}{R^2} = 1$ , and hence  $\sum_{p=1}^w d_p = 1$ . Accordingly, the importance of the explanatory variables can then be ordered by the index  $d_p$ . That is, the relative importance, as indicated by the proportion of the observed variance or  $R$ -square to which an explanatory variable contributes, is *relative* to the other explanatory variables included in a regression model.

Thomas et al. (1998) pointed out one caveat with the method of Pratt's importance measures. Pratt's importance measures occasionally produce negative importance values. This is a counterintuitive characteristic for importance interpretation (i.e., variance explained, which should be always positive). Pratt's measures are population-defined, but both  $\hat{\beta}_p$  and  $\hat{r}_p$  are sample estimates that are subject to sampling errors, hence small negative Pratt's measures could occur due to chance. As Pratt (1987) notes, a large negative index would suggest that the statistical importance is too complex to be captured in a single index (Thomas et al., 1998).

Pratt's important measures can be applied to the results of a latent variable regression even when the outcome variable is a latent variable in a SEM as specified in Fig. 16.1. This application will be explained and demonstrated in the next section with an example of investigating the importance of three different types of test-taking strategy to the latent test performance in a reading comprehension test. In doing so, we can shed light on the roles of test-taking strategies in test-takers' performance and this will provide useful information for test score validation.

## Illustration

This section provides an illustration of the latent variable regression in SEM, as specified in Fig. 16.1, with Pratt's measures to explain the score variation in the latent variable by the three types of test-taking strategy.

### *Participants*

The data were collected in 2012 as part of a series of validation studies undertaken during the pilot test for the revised CELPIP-G reading test. A total of 189 participants from a wide range of cultural/language backgrounds and English language proficiency levels were recruited from the greater Vancouver area in Canada. Most of the participants were from ESL and new immigrant communities, but some were from mainstream workplaces. Participants-reported time of having lived in Canada ranged from .08 to 54 years with a median of 1.33. About 66% of the participants were females, and participants' age ranged from 18 to 70 with a mean of 35.9 years. As for employment, about 24.3% were employed (part- or full-time), 31.3% were unemployed or not looking for a job, 15.7% were students (part- or full-time) obtaining a formal degree or diploma, and 28.6% were taking part-time ESL programs.

### *Measures*

**The CELPIP-G Reading Pilot Test** The CELPIP-General is a computerized assessment of functional English language proficiency. The construct of functional reading proficiency in English is defined as an individual's ability to engage with, understand, interpret, and make use of written English texts to achieve day-to-day and general workplace communicative functions. The official test is approved by the Canadian federal government as a screen for immigration. Test-takers include both English-first and ESL speakers. A paper-and-pencil version of the CELPIP-G reading pilot test was administered. The pilot test form included the four task types (testlets) of the standard operational test. Each task type required the test-takers to read a passage and then answer a set of questions that assessed their comprehension associated with that passage.

The four task types vary in genre, number of items, linguistic/cognitive skills required, and time allowed to complete. The tasks become increasingly difficult as the test progresses.

The first task *Reading Letter Correspondences* focuses on a social or workplace email correspondence. This task requires an understanding of and an ability to make inferences from common sociolinguistic and pragmatic aspects of written

correspondence, a good understanding of high frequency vocabulary and the ability to make sense of common syntactic structures. The second task *Reading to Understand & Apply Diagram Information* focuses on graphically presented information and an email cloze section that requires integration of the graphically presented information with the email communication. The third task *Reading for Factual Information* presents factual information (in a genre similar to a Wikipedia entry). This task requires test-takers to identify, extract, and paraphrase information in the texts. The fourth task *Reading for Viewpoints* presents more complex and abstract ideas in an editorial genre. Test-takers need to understand and contrast complex, abstract ideas.

Table 16.1 provides a summary of the test structure, reliability estimates, and descriptive statistics at the task level (in proportion of correct answers). The increasing task difficulties were observed by the mean proportion correct scores of 0.76, 0.66, 0.59, and 0.42 from task types one to four. There are 38 questions in the entire test. All items are in a multiple-choice response format scored as correct = 1 and incorrect = 0.

**Cognitive Test-Taking Strategies Survey** Ten test-taking strategies were investigated in this study. These strategies were selected from the qualitative research review. The strategies were selected based on the degree that previous qualitative research considered them to be relevant to test-taker response behavior (Cohen & Upton, 2006, 2007). Each strategy was presented as a statement beginning with the phrase “When answering a question...” and followed by the description of the strategy. Strategies S1–S3 were chosen to reflect test-takers’ engagement in comprehending-meaning (CM; e.g., I needed to understand the main ideas of the passage), S4–S7 to reflect test-takers’ engagement in score-maximization (SM; e.g., I tried to guess from other sentences), and S8–S10 to reflect test-takers’ engagement in test-wiseness (TW; e.g., I simply chose the answer that seemed the least wrong). Immediately after completing each task, test-takers reported the extent to which they had been engaging in each strategy on a 5-point scale of 1 = *never*, 2 = *seldom*, 3 = *sometimes*, 4 = *often*, and 5 = *always* by reflecting on their own cognitive engagement in reading the passage and answering the questions within the task. In order to make sure that low ability test-takers understood the meaning of the statements, and that all test-takers interpreted the statements in the same way, the researchers explained the meaning of each statement carefully to the test-takers before the data collection. Appendix A lists the individual strategy survey items and reports and descriptive statistics.

The last two columns of Table 16.2 report the descriptive statistics for each strategy type based on the mean collapsed over the item scores corresponding to each strategy type. It can be seen that test-takers reported engaging in the comprehending-text strategy fairly frequently (3.56–4.00) and the endorsement increased as task difficulty increased from tasks one to four. Test-takers reported using score-maximization skill relatively less (2.55–2.86), the extent to which they use this strategy varied across tasks but did not necessarily show a clear correspondence with task difficulty. Test-takers reported using test-wiseness strategy even less

**Table 16.2** Correlations among and descriptives of test-taking strategy types– by tasks

	CM	SM	TW	M	SD
Task-1 letter					
CM	1.00			3.56	0.74
SM	0.39	1.00		2.66	0.91
TW	0.16	0.60	1.00	2.11	0.86
Task-2 diagram					
CM	1.00			3.69	0.80
SM	0.32	1.00		2.55	0.84
TW	0.22	0.65	1.00	2.16	0.89
Task-3 information					
CM	1.00			3.84	0.79
SM	0.33	1.00		2.68	0.95
TW	0.23	0.68	1.00	2.38	1.06
Task-4 views					
CM	1.00			4.00	0.81
SM	0.32	1.00		2.86	0.91
TW	0.17	0.57	1.00	2.73	1.09

Note. *CM* comprehending-meaning, *SM* score-maximizing, *TW* test-wisness

(2.11–2.73) and the endorsement increased as task difficulty increased. The correlations among the three types of strategy are shown on the left side of Table 16.2. The patterns of correlations are very similar across the four tasks. There was a medium level of correlation between the report of using comprehending-meaning and score-maximizing strategies, a low correlation between comprehending-meaning and test-wisness strategies, and a high correlation between score-maximization and test-wisness strategies. All correlations were statistically significant.

### ***Analysis – Latent Variable Regression with Pratt’s Important Measures***

With the first task of Letter (reading correspondences) as an example, Fig. 16.1 depicts the latent variable regression using the SEM convention of a path diagram. The outcome variable is a latent variable with 11 observed item scores (indicator variables) associated with the passage of the first task. Because the observed item scores are binary variables, a tetrachoric correlation matrix will be analyzed. The three correlated explanatory variables are the scores of the three types of test-taking strategy being averaged over the corresponding items of each type. The same SEM will be specified for each of the four tasks separately. The analyses were all conducted for each of the four tasks separately in *Mplus* 7.4 (Muthén & Muthén, 1998–2015) using the default estimator of WLSMV (weighted least squares with means and variances for categorical variables). For each of the 4 tasks (outcome variables),

Pratt's importance measures based on the latent variable multiple regression were computed for each predictor variable and then compared across the tasks. The R-square values were partitioned using the same formula in (2), the only difference in this application is that the R-square value is for a latent outcome variable in a SEM rather than an observed outcome variable as in the case of a multiple regression. Thus, we can compute the importance of the three types of test-taking strategy to the explained variance of the latent task performance. Appendix B provides the syntax for the specified SEM model as well as the instructions for how to obtain the  $\beta$  and  $\hat{r}$  needed to compute the Pratt's importance measures using *Mplus*.

## Results

Table 16.3 reports the fit indices for each of the four latent variable regressions (for the four tasks) for the specified SEM model shown in Fig. 16.1. The models yielded good fit to the data for all four tasks. All CFI and TLI were greater than 0.95 (except for Task-1 = 0.92). RMSEA were all less than 0.05 and WRMR were all less than 1.00.

Table 16.4 reports the results of the latent variable multiple regression and the Pratt's measures broken down by the four tasks. The modeled  $R^2$  in the first column shows that about 30%–37% of the variation in the latent performance was accounted for by the three types of strategy. The next two columns report the standardized partial regression coefficient  $\beta$  of each strategy type and its corresponding  $p$ -value. The findings show that, all four tasks, test-takers' self-report of engagement in comprehending text meaning was not found to be a statistically significant explanatory variable for latent performance. As for self-report of using score-maximizing skill, it was found to significantly but negatively associated with the latent performance on the two easier tasks (Task-1 and Task-2), but was not a statistically significant explanatory variable for the two more difficult tasks (Task-3 and Task-4). Finally, self-report of test-wiseness was found to be negatively associated with the latent performance for all tasks.

The column  $\hat{r}$  in Table 16.4 reports the estimated simple correlation between a strategy type and the latent performance. The product of  $\beta\hat{r}$  for a specific explanatory variable is the unstandardized Pratt's importance measure and shows the proportion of variation that was accounted for by that explanatory variable. The sum of the

**Table 16.3** SEM fit indices of the latent task performance regressed on the strategy types— by tasks

	Task-1 letter	Task- 2 diagram	Task-3 information	Task-4 views
CFI	0.920	1.000	0.996	0.968
TLI	0.922	1.000	0.996	0.969
RMSEA	0.045	0.000	0.014	0.030
WRMR	0.907	0.764	0.745	0.818

Note. *CFI* comparative fit index, *TLI* Tucker & Lewis index, *RMSEA* root mean square error of approximation, *WRMR* weighted root mean square residual

**Table 16.4** Results of latent variable multiple regression and Pratt's importance measures

Task-1 letter	$p$	$\hat{\beta}$	$\hat{r}$	$c$	$d$
CM	0.090	0.217	-0.021	-0.005	-0.014
SM	<b>0.007</b>	<b>-0.365</b>	-0.440	0.161	0.479
TW	<b>0.001</b>	<b>-0.393</b>	-0.456	0.179	0.535
Modeled $R^2 = 0.329$			Sum	0.335	1.000
Task-2 diagram	$p$	$\hat{\beta}$	$\hat{r}$	$\hat{\beta}\hat{r}$	$d$
CM	0.165	0.171	-0.037	-0.006	-0.019
SM	<b>0.039</b>	<b>-0.350</b>	-0.437	0.153	0.448
TW	<b>0.011</b>	<b>-0.406</b>	-0.479	0.194	0.570
Modeled $R^2 = 0.330$			Sum	0.341	1.000
Task-3 information	$p$	$\hat{\beta}$	$\hat{r}$	$\hat{\beta}\hat{r}$	$d$
CM	0.143	0.146	-0.036	-0.005	-0.014
SM	0.092	-0.194	-0.455	0.088	0.231
TW	<b>&lt;0.001</b>	<b>-0.464</b>	-0.645	0.299	0.783
Modeled $R^2 = 0.373$			Sum	0.382	1.000
Task-4 views	$p$	$\hat{\beta}$	$\hat{r}$	$\hat{\beta}\hat{r}$	$d$
CM	0.518	0.061	-0.038	-0.002	-0.008
SM	0.623	-0.052	-0.309	0.016	0.053
TW	<b>&lt;0.001</b>	<b>-0.487</b>	-0.600	0.292	0.626
Modeled $R^2 = 0.301$			Sum	0.306	1.000

Note. Statistically significant predictions are highlighted in bold face

three  $\hat{\beta}\hat{r}$  over the three explanatory variables indicates the R-square values obtained from the Pratt's method. For all four tasks, the R-square values based on the sum of  $\hat{\beta}\hat{r}$  are all very close to R-square values reported by *Mplus* (only different in the second or third decimal place). The last column of  $d$  shows the standardized Pratt's importance measures (standardized by the R-square values from the Pratt's method), which should add up to one.

For all four tasks, the Pratt's importance measure  $d$  reveals that test-takers' self-report of engaging in comprehending text meaning did not contribute to any of the R-squares (close to zero, small negative of -0.014, -0.019, -0.014, -0.008 due to chance). As for self-report of employing test-wiseness strategy, it turned out that, for all four tasks, reporting *not* using the test-wiseness strategy (a negative predictive relationship shown by  $\hat{\beta}$ ) was the most important contributor to the R-squares. Its contribution (through a negative prediction) became more prominent as task difficulty increased from Task-1 to Task-4 (cf., 0.535, 0.570, 0.783, and 0.955 of the R-square, respectively). Finally, *not* using the score-maximization strategy turned out to be a fairly important contributor to the R-squares of the easier tasks (0.535 and 0.448 of the R-square for Task-1 and Task-2), but became increasingly less important as task difficulty increased (0.231 and 0.053 of R-square for Task-3 and Task-4).



## Discussion

Task-specific descriptive statistics for the report of strategy use show that the extent to which test-takers employ test-taking strategies varies with tasks. This suggests that test-taking strategies should be conceptualized as a state of process that function as a reaction to different tasks employed by individuals with different levels of ability, rather than a trait-like construct that can be observed in a static manner.

The use of a latent variable as the outcome variable circumvents the problems of measurement errors in a regression. The method of Pratt's importance measures further helps to transform the regression results into new information in term of score variation accounted for by each explanatory variable. These measures have a very straight-forward interpretation and provide very useful information about the importance of the explanatory variables.

Findings from latent variable regression suggest that test-takers' reports of engagement in comprehending text meaning does not necessarily lead to better performance nor does report of low level of engagement lead to poorer performance. This result is common to all four tasks. This finding may seem surprising at first but is more plausible after some consideration. When individuals read in a very intense situation such as in a setting of testing, they will engage in the ability that is being tested, a.k.a., comprehending the text meaning. This is manifested in the present sample. However, an individual's report of efforts in engaging in the measured ability does not necessarily guarantee a consequent comprehension and may lead to both successful and unsuccessful outcomes. Following this reasoning, it is not surprising that self-report of engagement in the comprehending-meaning strategy is a poor explanatory variable of the actual outcomes.

Test-takers' reports of using score-maximization skills was found to be associated with the test performance only when the task is relatively easy, but not the case when the task is difficult. When tackling Task-4, the most difficult task, test-takers' reports of using score-maximization strategy were almost fruitless in affecting performance (only 1.6% of the variation in the latent variable). This finding suggests that if fluency in comprehension (reading with automaticity) is the construct to be measured, score-maximization skills based on fragmented and partial understanding may be considered as a construct-irrelevant source of score variation.

Report of *not* using test-wiseness turns out to be most related to score variation compared to the other two types of strategy. The relationship becomes stronger as task difficulty increases. The level of (not) using test-wiseness, if reported with high credibility, can inform test-takers' ability when conditioning on the tasks' difficulty (or tasks' difficulty when conditioning on the test-takers' ability).

We believe that evidence based on this conditional information about SM and TW strategy use, although less straightforward, can provide a strong argument for validity. That is, this evidence supports a claim that test-takers who utilize SM and TW strategies do so when they are at the certain limits of their ability level. If this finding is related to the test scores, then it can be inferred that the scores are indeed related to the target construct, in this case, the ability to comprehend and use English written texts.

In this chapter, we provided an alternative approach to understanding test score variation through understanding one aspect of the processes of responding- test-taking strategies. Readers are reminded that in this study, test-takers' reports of strategy use is a reflection of their agentive processes in responding to the written texts and questions, surveyed through a limited number of pre-specified strategies. It is, at best, a partial form of meta-cognition and by no means reveals the true cognitive processes of taking a reading comprehension test. However, as Storey (1997) pointed out, the validation of tests involving cognitive processes will not be complete unless it includes some examination of the processes by which solutions to test tasks are actually reached. Moreover, no matter how these processes are understood or further analyzed through cognitive or even neural research, the findings provided herein nonetheless shed light on how test-takers' engagement with a test inform us of test-takers' ability.

We also would like to point out that a process-based approach to validation is not a replacement for the other three sources of validity evidence (content, internal structure, and relations to other variables) as outlined in *the Standards for Educational and Psychological Testing*. Rather, we see validation as an ongoing practice of social science. There is no concrete determination of validity based solely on one piece of evidence. Our approach and findings only offer an approach to addressing one gap in validity research. Many others remain to be filled.

## Appendices

### Appendix A

Statements of the test-taking strategy survey items and descriptive statistics

Type	Item #	When answering a question, ...	M	SD
CM	1	I needed to understand the main ideas of the passage.	3.81	1.07
	2	I needed to understand some specific sentences in the passage.	3.73	1.00
	3	I needed to read some parts again carefully.	3.75	0.94
SM	4	I quickly summarized or took notes.	2.42	1.20
	5	I translated some words/sentences of the passage.	2.32	1.20
	6	I tried to guess from other sentences.	3.19	1.18
	7	I used clues in the other questions to guess the answer.	2.84	1.22
TW	8	I simply chose the answer that seemed the least wrong.	2.55	1.19
	9	I selected an option that had an important word.	2.49	1.16
	10	I guessed blindly.	2.01	1.19

Note. *CM* comprehending-meaning, *SM* score-maximizing, *TW* test-wiseness. The descriptive statistics were computed based on the responses to all four tasks

## Appendix B

1. Note that texts in italic face followed by “!” are the descriptions of the Mplus syntax.
2. Use Mplus standardized outputs under the heading of “STD” to obtain both the  $\hat{\beta}$  and  $\hat{r}$  for computing the Pratt’s measures.

-----

Title: Regression of the Latent Task Performance (Task-1, Letter) on the 3 the Observed Strategy Types

Data: File is Strategy for Mplus.dat;

Format is 455F5;

Variable:

NAMES ARE id L1–L11 LCM LSM LTW;

! L1-L11 are scores for reading questions of Letter (Task-1).

! LCM, LSM and LTW are the observed scores for CM, SM and TW strategies.

USEVAR ARE L1-L11 LCM LSM LTW;

CATEGORICAL ARE L1-L11;

Missing are all (99);

Model:

P1 by L1-L11; ! Measurement model for latent performance (P1) for Letter (Task-1)

P1 on LCM LSM LTW; ! Latent variable regression (P1) on three types of Strategy  
! To obtain correlations for P1 with LCM LSM and LTW, replace “on” by “with”  
(no “on” command in the model)

Output: Standardized; ! To obtain  $R^2$ ,  $\hat{\beta}$  and  $\hat{r}$  for computing Pratt’s measures  
(Use “STD” standardized)

## References

- Abbott, M. L. (2006). ESL reading strategies: Differences in Arabic and Mandarin speaker test performance. *Language Learning, 56*, 633–670.
- Allen, S. (2003). An analytic comparison of three models of reading strategy instruction. *International Review of Applied Linguistics in Language Teaching, 41*, 319–338.
- 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.
- Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*, 41–66.
- Cohen, A. D. (1984). On taking language tests what the students report. *Language Testing, 1*, 70–81.
- Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly, 3*, 307–331.

- Cohen, A. D. (2012a). Test-taking strategies. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoyhoff (Eds.), *The Cambridge guide to second language assessment* (pp. 96–104). Cambridge, UK: Cambridge University Press.
- Cohen, A. D. (2012b). Test taker strategies and test design. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing in a nutshell* (pp. 262–277). Abingdon, UK: Routledge.
- Cohen, A. D., & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks. In *TOEFL Monograph Series, MS-33*. Princeton, NJ: Educational Testing Services.
- Cohen, A. D., & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®. *Language Testing, 24*, 209–250.
- Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*, 209–226.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.
- Lee, J. Y. (2011). *Second language reading topic familiarity and test score: test-taking strategies for multiple-choice comprehension questions* (Doctoral dissertation). The University of Iowa, Iowa City, IA. Retrieved from <http://ir.uiowa.edu/cgi/viewcontent.cgi?article=2717&context=etd&unstamped=1>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.
- Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Authors.
- Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*, 199–215.
- Newton, P. E. (2016). Macro- and micro-validation: Beyond the 'five sources' framework for classifying validation evidence and analysis. *Practical Assessment, Research & Evaluation, 21*, 1–13. Available online: <http://pareonline.net/getvn.asp?v=21&n=12>
- Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In T. Pukilla & S. Duntaneu (Eds.), *Proceedings of second Tampere conference in statistics* (pp. 245–260). Tampere, Finland: University of Tampere.
- Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: A structural equation modelling approach. *Language Testing, 15*, 333–379.
- Roizen, M. A. (1984). *Test performance vis a vis test-taking strategies in reading comprehension of English as a second language*. ERIC Document Reproduction Service No. ED 224 350. Retrieved from <http://ir.uiowa.edu/cgi/viewcontent.cgi?article=2717&context=etd&unstamped=1>
- Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*, 441–474.
- Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing, 14*, 214–231.
- Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research, 45*, 253–275.
- Wu, A. D., & Stone, J. E. (2015). Validation through understanding test-taking strategies: An illustration with the CELPIP-General reading pilot test using structural equation modeling. *Journal of Psychoeducational Assessment, 34*, 1–18.
- Zumbo, B. D. (2005, July). Reflections on validity at the intersection of psychometrics, scaling, philosophy of inquiry, and language testing. *Samuel J. Messick Memorial Award Lecture*. Lecture provided at the 27th Language Testing Research Colloquium, Ottawa, Canada.
- Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Psychometrics* (Vol. 26, pp. 45–79). Amsterdam, The Netherlands: Elsevier Science B.V.

# Chapter 17

## An Investigation of Writing Processes Employed in Scenario-Based Assessment

Mo Zhang, Danjie Zou, Amery D. Wu, Paul Deane, and Chen Li

### An Investigation of the Writing Processes Employed in Scenario-Based Assessment

In this study, we demonstrate the use of keystroke logs in the context of assessment in terms of validating assessment design and scores. The evidence presented herein focuses on validity evidence in line with Messick's (1989, 1995) substantive validity, which focuses on evidence about the process of responding. Specifically, we investigated whether a particular assessment structure has an impact on students' writing processes executed in responding to an essay task. Each of the key presses in the writing process is recorded, and features extracted from the keystroke logs are analyzed in terms of their association with performance.

The specific assessment structure examined in this study is scenario-based with a theoretically determined task order, which can be arguably described to as “scaffolded.” The term “scaffolding” is not new in education. The literature on the use of scaffolding can be traced back to Lev S. Vygotsky (1896–1934), a Russian developmental psychologist who studied children's cognitive development and concluded that children can perform beyond their maturational level with high-quality guidance (e.g., Vygotsky, 1978). The concept was elaborated by Jerome S. Bruner (1915–2016) who coined the term, “scaffolding theory,” when studying the language acquisition of young children (Bruner, 1978). Scaffolding has since been

---

M. Zhang (✉) • P. Deane • C. Li  
MS T03, Educational Testing Service, Princeton, NJ, USA, 08541  
e-mail: [MZhang@ETS.org](mailto:MZhang@ETS.org); [PDeane@ETS.org](mailto:PDeane@ETS.org); [CLi@ets.org](mailto:CLi@ets.org)

D. Zou • A.D. Wu  
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: [Zoudanjie7@gmail.com](mailto:Zoudanjie7@gmail.com); [Amerywu@mail.ubc.ca](mailto:Amerywu@mail.ubc.ca)

studied extensively as a way to improve students' learning in the context of classroom instruction (e.g., Jackson, Krajcik, & Soloway, 1998; Puntambekar & Hubscher, 2005), in language acquisition (e.g., Greenfield & Smith, 1976), and in concept development (Pea, 2004).

With some exceptions, the literature on scaffolding in *educational assessment* is limited. For example, Shepard (2005) argued that formative assessment and scaffolding following Vygotsky's Zone of Proximal Development (ZPD) are fundamentally the same. That is, a central goal of formative assessment is to help students take ownership of their learning. Similarly, ZPD implies that experience in learning activities that were guided initially by a teacher can help the student build mental capacity, internalize the learning process, and achieve results that couldn't otherwise be produced.

Scaffolding can be thought of as an educational structure (Pea, 2004). The structure guides students' thinking *and* behavior, with the intention that students eventually internalize and re-produce the thinking and doing on their own. Scaffolding is, therefore, a means of modeling good thinking and doing in learning.

In the assessment context, the scaffolding structure can be extended to (a) encouraging and modeling good practice and (b) obtaining an evaluation of students' readiness to undertake tasks *within a targeted zone of proximal development*. These goals can be achieved through channeling and focusing, and modeling (Pea, 2004). "Channeling and focusing" refers to constraining the dimensions of the task to increase the likelihood of a learner's effective action, focusing a learner's attention on desired, relevant features of a task, and building a sense of directness of a learner's activities towards successful completion of the task. "Modeling" refers to exemplifying more sophisticated approaches to solving the problem. Every assessment provides some form of channeling and focusing in order to maximize the probability that student responses will address the intended construct. It is less common for assessments to be designed to model effective strategies (though see Wiggins, 1992). One way to accomplish this goal is developed in *scenario-based assessments*, a hybrid form of assessment in which individual items are designed to assess specific skills, but in which the sequence of tasks models the steps that an expert would take to solve a complex performance task (Deane et al., 2015; Sabatini, O'Reilly, Halderman, & Bruce, 2015; Sheehan & O'Reilly, 2012).

### ***Rationale for the ETS CBAL Scenario-Based Assessment***

One scenario-based assessment (SBA) design developed for the Cognitively-Based Assessment *of, for, and as* Learning (CBAL) research initiative is a summative writing assessment design focused on argumentative essays. It includes an organizing scenario that provides a purpose for writing and a sequence of tasks designed to accomplish that materials. Those tasks make use of source materials, move students through a sequence of lead-in questions designed to get students to think about and analyze the contents of the reading, after which students undertake a culminating essay task. The individual reading and writing tasks are designed to channel and

focus student work. The overall sequence of tasks is designed to model key steps an expert writer would undertake in an independent performance task.

Zhang, van Rijn, Deane, and Bennett (2016) summarized the design purposes of the lead-in items and tasks, which derive directly from notions of scaffolding. (See Bennett, Deane, and van Rijn, 2016, for a detailed description of the theory underlying this assessment design.) The lead-in tasks, *ordered in a particular theoretically determined way*, are designed to encourage students to engage with the sources, helping to reduce differences in topic familiarity; to measure the component skills required to write effectively; to model the processes employed in extended projects; and to step students through a short version of that process to prepare them better for undertaking a culminating performance task. The intention is that this structure (or task ordering) will provide a more meaningful depiction of student skills than would be obtained by having students write an essay without a mechanism to engage them with the source materials.

We would expect the depiction of skills through the lead-in task design to be more meaningful for at least two reasons. First is our belief that, under a traditional test design, many struggling students would be unlikely to complete the culminating essay performance. Our hope is that these students will be primed to attempt the essay as a result of engaging with the sources and completing the lead-in tasks. A second reason is that, in a complex task, performance depends on applying multiple sub-competencies simultaneously. If we administer the essay task without the lead-in questions, we are likely to identify the students who can coordinate all the necessary sub-competencies effectively. However, it is not clear whether we can differentiate those students who can almost complete the task (but gave up) from those who are not anywhere near as close to full competency. Further, we will have no information about the student's skill with respect to the sub-competencies themselves. Therefore, a task sequence that encourages less able students to attempt the essay may evoke more and better evidence than would a traditional design.

### ***Study Purpose***

In this exploratory study, we evaluated whether the SBA structure had an impact on the writing processes that students executed in responding to the essay task. Students' writing processes were recorded via keystroke logs, which are described in more detail next.<sup>1</sup> The research questions were: (1) *How do the process features that predict human essay scores differ in the scenario-based vs. the alternative condition?* (2) *How do students' writing processes differ in the scenario-based design as compared with an alternative design?*

---

<sup>1</sup>The keystrokes are recorded using ETS' patent-pending keystroke logging system.



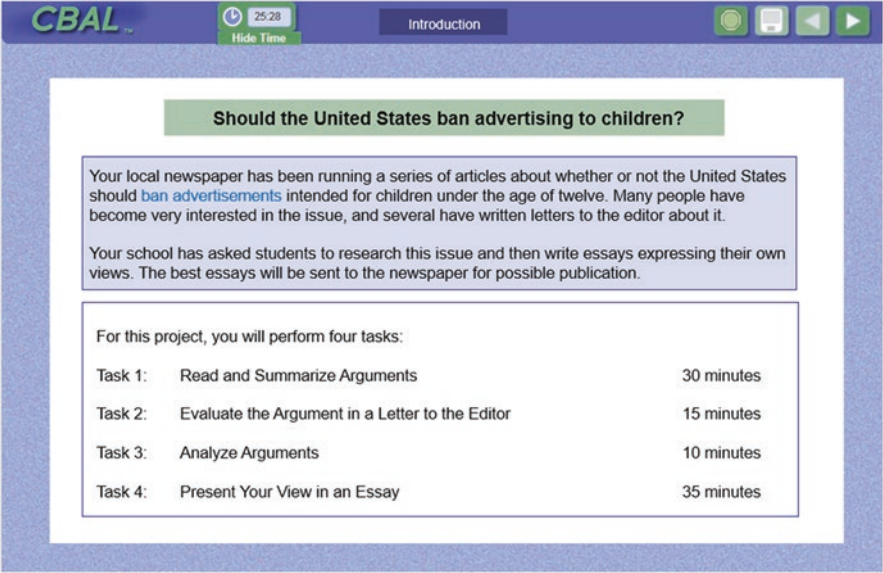
## ***Keystroke Logging***

Keystroke logging (KL) is a tool that records the mechanical and temporal characteristics of an individual text production process. A typical KL system identifies and records several kinds of information, including the type of action (e.g., an insertion, a deletion), the duration of an action (e.g., how long does the insertion last before transitioning to a different action?), the location of an action (e.g., does the insertion occur within a word?), and the time-point of an action in the composition process (e.g., does the insertion occur right before final submission?). Keystroke logging can be complemented by Natural Language Processing techniques, such as recognizing spelling errors and corrections, and identifying word types and linguistic complexity. To our knowledge, research on writing processes in the context of educational assessment is quite sparse (e.g., exceptions include Deane, 2014; Deane & Zhang, 2015; Guo, Deane, van Rijn, Zhang, & Bennett, 2016; Zhang, Hao, Li, & Deane, 2016). As the literature suggests, information recorded and extracted from KLs can provide rich evidence for inferring the cognitive processes underlying composition. The value of evaluating composition processes in summative and formative assessment could be considerable, including improving test security, giving teachers and students diagnostic feedback to improve writing practice, characterizing differences among subpopulations beyond the quality of the final product, and providing validity evidence related to an assessment (e.g., Stone, 2016), as in the current study.

## **Method**

### ***Instrument***

The instruments for this study were developed through the CBAL initiative at ETS and focused on the assessment of argumentative writing. In the original CBAL assessment form (referred to as Original Form hereafter), students were told at the beginning of the test that they would be expected to compose an essay on a certain topic and were also informed that they would first respond to tasks intended to help them understand the topic. In one task, the students were presented with two readings and asked to evaluate hypothetical students' summaries of those readings, and then to write their own summary. In the next task, students were asked to analyze arguments presented in the readings, such as selecting claims that support or weaken an argument. Third, the students were asked to critique a hypothetical letter to the editor (a third reading selection) on the same subject. As described earlier in the chapter, these lead-in tasks were intended to help scaffold students' thinking on what is an argument, what is a claim, and how to formulate their own written argument. At the end of the assessment, the students were asked to write an argumentative essay on the same topic.



**Should the United States ban advertising to children?**

Your local newspaper has been running a series of articles about whether or not the United States should **ban advertisements** intended for children under the age of twelve. Many people have become very interested in the issue, and several have written letters to the editor about it.

Your school has asked students to research this issue and then write essays expressing their own views. The best essays will be sent to the newspaper for possible publication.

For this project, you will perform four tasks:

Task 1:	Read and Summarize Arguments	30 minutes
Task 2:	Evaluate the Argument in a Letter to the Editor	15 minutes
Task 3:	Analyze Arguments	10 minutes
Task 4:	Present Your View in an Essay	35 minutes

**Fig. 17.1** The original test form

The Original Form was an established CBAL assessment, “Ban Ads,” (Fig. 17.1). Its topic concerns whether advertisements directed at children should be banned.

The alternative form analyzed in this study was the reverse ordering of the Original Form, which breaks the scaffolding or the task-ordering effect (referred to as Reversed Form hereafter). By comparing the Original Form with the Reversed Form, we can analyze the effect of writing the essay with and without the topical preparation afforded by having the lead-in tasks come first *and* with and without the argumentation scaffolding provided by those tasks (i.e., summarize, evaluate, and critique arguments).

Keystroke logs were recorded for all essay submissions, though there was a small loss of logs due to unexpected technical glitches. For each valid keystroke log, we extracted 29 writing process features based on Zhang and Deane (2015). The list of the features and their substantive descriptions can be found in the aforementioned article, as well as in the Data Analyses section later in this chapter.

## *Participants*

This study used data collected from a larger experimental study that focused on scenario-based assessment in argumentative writing (Zhang, van Rijn et al., 2016). Data were collected over a one-month period, from September 29, 2014 to October 31, 2014. In that experiment, an English language arts (ELA) scenario-based assessment form was administered, along with three alternative forms. The four forms

were randomly distributed among students within classrooms. Students were randomly assigned to one of four test forms within each participating class. Each form was administered in two 45-min sessions taken close together in time but on different days. The complete original data set consists of a convenience sample of 1082 eighth-grade students from various states in the US. The analyses conducted in this study used the data collected using two of those four forms, which allowed us to compare writing processes with and without the preparation afforded by the scaffolded, scenario-based design.

In data cleaning, 171 cases from the two selected test forms were deleted (cleaning procedures described next), leaving 444 students for analysis in this study. Following Deane and Zhang (2015) and Zhang and Deane (2015), the following procedures were employed. First, we eliminated cases for which keystrokes were incorrectly collected or recovered. Second, a few students were given accommodations when taking end-of-year assessments and their responses were excluded. Third, records with essay scores of 0 were eliminated as this coding indicates an aberrant response. Fourth, cases with fewer than 25 words were filtered out and, for similar reasons, cases with total writing times of less than 3 min were excluded. Finally, responses with overly long writing time (35 min) were removed. The final retained sample consisted of 444 essay responses.

The final sample of 444 students taking the Original and Reversed Forms consisted of 52% female and 48% male students. The large majority of students were reported to be Caucasian (83%), followed by African American (9%), Hispanic (3%), and Asian (3%), with the remaining 2% reported as mixed race, Native American, Hawaiian/Pacific islanders, and Middle Eastern. About 25% of students were indicated as participating in the free or reduced price lunch program, which is an index of Socioeconomic Status (SES). Demographic information was not available for eight students.

In the typical CBAL writing assessment, human raters apply two scoring rubrics to grade essay performance, one generic rubric (Rubric 1, shown in Appendix A) for evaluating basic writing skills (e.g., grammar, mechanics) and the other genre-specific (Rubric 2, shown in Appendix B) for evaluating higher-order skills specific to the task (in this case, argumentation skills, e.g., quality of arguments and appropriateness of evidence). The mean scores for the Original Form were 2.17 ( $SD = 0.89$ ) and 2.55 ( $SD = 0.97$ ) on Rubric 1 and Rubric 2, respectively. The comparable values for the Reversed Form were 2.32 ( $SD = 0.93$ ) and 2.60 ( $SD = 1.05$ ). The means on the two forms were not statistically significantly different on either rubric. This comparison is also reported in Zhang, van Rijn et al. (2016) where students taking the two forms received comparable essay and test scores even though students taking the scaffolded form spent less time on the task and wrote shorter responses.

## Data Analyses

For the purpose of understanding which writing process features predicted essay score, we applied multiple linear regression analyses. The regression analyses were conducted separately by form. Human rating was the dependent variable, and the keystroke features were used as predictors. All 29 keystroke variables from Zhang and Deane (2015) were entered into the regression model. The assumptions and diagnostics for linear multiple regression, such as linearity, normality, homoscedasticity, multicollinearity and so on, were checked. Results indicated no severe issues.

Pratt's (1987) importance measures were calculated based on the regression results as an aid in interpreting differences across forms. Pratt's measure for the unique importance of the  $p^{\text{th}}$  predictor is expressed as the product of its standardized regression coefficient  $\hat{\beta}_p$  and its simple correlation  $r_p$  with the outcome variable. The  $R^2$ -standardized Pratt's measure,  $d_p$ , was later described in Thomas, Hughes, and Zumbo (1998) as:  $\sum_{p=1}^w \hat{\beta}_p \hat{r}_p = R^2$ . The statistic follows that,  $\sum_{p=1}^w \frac{\hat{\beta}_p \hat{r}_p}{R^2} = 1$ , hence  $\sum_{p=1}^w d_p = 1$ . Accordingly, the importance of the predictors can then be ordered by  $d_p$ , the proportion of the  $R^2$  explained by the  $p^{\text{th}}$  predictor.<sup>2</sup>

We applied this method to additively attribute the regression  $R$ -square to the each of the 29 predictors. The relative importance of each the 29 predictors was then ordered by the size of Pratt's measures. In addition, because of the additive property, the sum of the Pratt's measures over a set of the predictors, indicates the joint importance of that set of predictors. Of note is that Zhang and Deane also demonstrated that 21 of the 29 features could be classified as belonging to one of four dimensions. By summing over the keystroke features associated with each one of the four dimensions reported in Zhang and Deane (2015), we obtained the joint importance measures aggregated at the dimension level. Then we compared the Original and Reversed Forms based on the additive predictive value of those features. The dimensions associated with those keystroke features, along with descriptions of the features are given in Table 17.1 (adapted from Zhang and Deane 2015).

To further examine where the differences might occur we ran multivariate analyses of variance on the process features for the two test forms, and compared the mean feature scores between the two forms. We also compute Cohen's  $d$  values comparing the two feature means to obtain an idea of the practical significance of the differences.

---

<sup>2</sup>Researchers had been using these measures like these for many years (e.g., Carlson, 2014; Chase, 1960) but these measures lacked a proper rationale until the axiomatic justification by Pratt. As noted by Pratt's himself, negative measures can occur. Small negative values could be due to capitalization on chance because both  $\hat{\tau}_p$  and  $\hat{\beta}_p$  are sample estimates. Large negative Pratt's measures can reflect a suppression effect or multicollinearity (Budescu, 1993; Thomas, Hughes, & Zumbo, 1998).

**Table 17.1** Dimensions and process features

Dimension	Process feature	Description
Local and word level editing	CorrectedTypo	Extent to which correction of mistyped words occurs.
	LongJump	Extent to which jumps to different areas in the text occur.
	MinorEdit	Extent to which words are edited to make only minor corrections.
	UncorrectedSpelling	Extent to which a spelling error occurs that is not corrected before another unrelated action is taken.
	WordChoice	Extent to which words are edited to produce completely different words, possibly suggesting deliberation about word choice.
Fluency	InterkeyPause	Extent to which pauses occur between keystrokes, suggesting general typing fluency.
	WordFinalPause	Extent to which pauses occur just before typing the last character in a word, a measure of general typing fluency.
	WordInitialPause	Extent to which pauses occur just after typing the first character in a word, which could reflect on the one hand, general typing fluency, or on the other hand, deliberation for word choice, retrieving spelling, or planning keystroke sequences for a word.
	WordInternalPause	Extent to which pauses occur within words during text production.
	WordSpacePause	Extent to which pauses occur in between words, suggesting general typing fluency.

(continued)

**Table 17.1** (continued)

Dimension	Process feature	Description
Phrasal and Chunk Level Editing	CharsInMultiWord Deletion	Extent to which deletion of characters occurs in the process of deleting multiple words.
	DeletedCharacter	Extent to which deletion of characters occurs.
	MultiWordDeletion	Extent to which multi-word text deletion occurs.
	MultiWordEditTime	Extent of time spent in deleting multiple words.
	NewContentChunk	Extent to which deleted text is replaced with edited text with new content.
Planning and Deliberation	EndSentencePunctuationPause	Extent to which pauses occur at the juncture of sentences, which may indicate planning and deliberation.
	EventsAfterLastCharacter	Extent to which the writing process reflected sequential drafting at the end of the text, which can be viewed (negatively) as a measure of the extent to which editing of any kind occurs.
	InSentence PunctuationPause	Extent to which pauses occur at a sentence-internal punctuation mark, which may reflect sentence-level planning.
	StartTime	Extent to which a pause occurs prior to beginning writing, possibly reflecting planning and deliberation.
	TimeSpentAtPhrasalBurst	Extent to which pauses occur at the beginning of a string of fluent text production, possibly suggesting planning and deliberation.
	TimeSpentBetweenPhrasalBurst	Extent to which pauses occur between strings of fluent text production, possibly suggesting conceptual planning and deliberation.

## Results

### *Feature Importance in Predicting Rubric Scores*

The process features extracted are more aligned to the writing basic skills as measured in Rubric 1. Using the Rubric 1 score as the outcome variable, the  $R$ -square for the multiple linear regression was 0.36 for the Original Form, and 0.38 for the Reversed Form. The comparable  $R$ -square values were 0.27 and 0.32 for the Original and Reversed Forms when Rubric 2 score was the outcome variable. Since the above results indicate that the Reversed Form calls more upon writing fluency than the Original Form does, it is not surprising that the  $R$ -squares for the Reversed Form are higher than the ones for the Original Form.

Table 17.2 provides the aggregated relative importance of each writing process dimension based on the Pratt's statistic. For the Original Form, the reliable variance in the Rubric 1 scores were predominantly explained by the features in the local and word level editing dimension (33%), followed by ones in the fluency dimension (29%). A different pattern is observed for the Reversed Form, for which half of the reliable variance in the Rubric 1 score variance was explained by the features in the fluency dimension (50%). The results also suggest that, with regard to Rubric 1, more reliable variance was accounted for by features in the planning and deliberation dimension for the Original Form (21%) than for the Reversed Form (14%).

These findings are interpretable within the context of the CBAL ELA writing assessment design. For the Original Form, we would suggest that because students were given the opportunity to familiarize themselves with the topic before having to write the essay, their essay performance was less dependent on how quickly they could generate text and get words onto the screen (as indicated via the fluency dimension), than about the extent to which they were able to plan and then revise and edit what they had written. For the Reversed Form, since the students were presented with the essay task without any extended preparation, the performance level appears to be overwhelmingly driven by general writing fluency.

**Table 17.2** Reliable variance in Rubric scores explained by Keystroke dimensions

Dimension (n of KL features)	Rubric 1 (basic writing skills)		Rubric 2 (argumentative writing skills)	
	Original	Reversed	Original	Reversed
Fluency (5)	0.29	0.50	0.24	0.46
Local and word level editing (5)	0.33	0.18	0.25	0.21
Phrasal and chunk level editing (5)	0.08	0.10	0.19	0.12
Planning and deliberation (6)	0.21	0.14	0.14	0.20
Remaining (8)	0.09	0.08	0.18	0.01

*Note.* Rubrics 1 and 2 are given in the appendices. Features in each dimension are given in Table 17.1



As for Rubric 2, the results are generally similar to those for Rubric 1. For the Original Form, the reliable variance in the rubric scores is mostly accounted by the features in the two editing dimensions (44% combined), whereas for the Reversed Form the reliable variance in the scores is largely accounted by the features in the fluency dimension (46% alone). The relatively strong association between the two rubrics may play a role in this observation. The Pearson correlations between Rubric 1 and Rubric 2 are 0.62 for the Original Form and 0.68 for the Reversed Form.

However, we note that, for both the test forms, the features in the phrasal and chunk level editing dimension appear to explain more reliable variance in essay scores in Rubric 2 than Rubric 1 (i.e., 19% vs. 8% for the Original Form; 12% vs. 10% for the Reversed Form), and the features in the fluency dimension appears to explain marginally more reliable variance in the essay scores in Rubric 1 than in Rubric 2 (i.e., 29% vs. 24% for the Original Form; 50% vs. 46% for the Reversed Form). This result is also consistent with our understanding of the keystroke features/dimensions in line with the purpose of each scoring rubric; that is, Rubric 1 intends to focus on the fundamental skills of writing (e.g., mechanics, grammar), which is more in tune with the fluency measured by keystroke features, whereas Rubric 2 is designed to evaluate genre-specific writing skills and high-level writing skills, which is aligned to the more global level editing and planning.

One final notable finding is the 18% of the remaining reliable variance that is not explained by the features from the four keystroke dimensions for the Original Form in Rubric 2. It is possible, but by no means certain, that students taking the Original Form have more cognitive resources available to pay attention to the quality of their writing by having to expend less effort on the lower-level writing aspects primarily measured in Rubric 1.

### ***Distinguishable Keystroke Features***

Next, we examined where the differences occur in the writing processes between the two test forms by looking into individual keystroke features. The MANOVA analyses revealed five features that were statistically significantly different between the two test forms, shown in Table 17.3. Four of these five features related to planning and deliberation, and one concerned phrasal and chunk-level editing.

Of interest is that the differences for all five features were statistically significantly different across the two forms (the practical significance of the between-form differences was small,  $|r| \sim .20$  to  $1.50$ ). Students in the Original Form condition planned and deliberated *less* immediately before beginning to type their essay and as they composed it. (The feature “StartTime” is coded such that longer pause duration prior to beginning writing receives lower values.) This result is most likely due to the assessment design; that is, in the Original Form, students did not necessarily need to conduct the same level of essay *real-time* planning as they did when given the essay task cold because they have had the benefit of the scaffolded topical and argumentative preparation provided by the lead-in tasks. In contrast, students taking

the Reversed Form had to learn about the topic, get familiar with the writing genre, and plan their essay without any structure or guidance before jumping into essay composition. They spent more time immediately before beginning to type and in planning and deliberation as they wrote.

Table 17.4 shows the correlations of each of the five features with essay score for each Form and Rubric. Of particular relevance to the above discussion are the results for the feature, “StartTime,” a measure of “pre-writing time” expended by the student, where pre-writing time is the duration between being presented with the essay composition screen and the first keystroke. (The correlations are negative because

**Table 17.3** Distinguishable features between original and revised forms

Keystroke feature	Dimension	Original mean (SD)	Reversed mean (SD)	F	Pr > F	d
StartTime	Planning & deliberation	0.26 (0.12)	0.21 (0.11)	15.09	0.0001	0.37
InSentencePunctuationPause	Planning & deliberation	1.47 (1.03)	1.81 (0.98)	12.75	0.0004	-0.34
TimeSpentBetweenPhrasalBurst	Planning & deliberation	0.29 (0.12)	0.33 (0.12)	9.97	0.0017	-0.30
EndSentencePunctuationPause	Planning & deliberation	1.24 (0.43)	1.36 (0.41)	8.42	0.0039	-0.28
CharsInMultiWordDeletion	Phrasal & chunk level editing	0.09 (0.11)	0.12 (0.15)	4.21	0.0407	-0.19

Note. “d” denotes Cohens’ d (Original Form minus Reversed Form). StartTime is coded such that longer time durations get lower values

**Table 17.4** Correlations with Rubric scores for selected features

Keystroke feature	Rubric 1 (basic writing skills)		Diff.	Rubric 2 (argumentative writing skills)		Diff.
	Original	Reversed		Original	Reversed	
StartTime	-0.13 ***	-0.20**	p = 0.218	-0.08	-0.23 ***	p = 0.054
InSentencePunctuationPause	0.18 **	0.16*	p = 0.431	0.07	0.13 *	p = 0.249
TimeSpentBetweenPhrasalBurst	0.01	0.11	p = 0.134	0.06	0.04	p = 0.406
EndSentencePunctuationPause	0.26 ***	0.11	p = 0.056	0.22 ***	0.17 *	p = 0.280
CharsInMultiWordDeletion	0.15 *	0.06	p = 0.151	0.13	0.03	p = 0.147

Note. \*\*\* p < 0.0001, \*\* p < 0.001, \* p < 0.05. Diff. refers to the probability that differences between the two correlations of Original and Reversed forms for a rubric would be obtained by chance alone. StartTime is coded such that longer time durations get lower values. Rubrics 1 and 2 are given in the appendices

the feature was coded such that longer pre-writing pause times were given lower values.) For Rubric 2, which measures the quality of ideas and evidence, on the Original Form, the correlation with score is not significantly different from zero, meaning that deliberating just before beginning to write has no relation to score, probably because that deliberation was done while working through the lead-in tasks. In contrast, the correlation of StartTime with essay score for students taking the Reversed Form was statistically significantly different from zero (though it missed being statistically significantly different from the correlation for those taking the Original Form; see Table 17.4). This result again suggests the possibility that the Reversed Form students needed to spend more time organizing their thoughts before beginning to compose.

## Discussion

In this chapter we report on a study in a program of research that informs test validation by focusing on evidence about the process of responding – that is, Messick's (1989, 1995) substantive validity evidence. We addressed the question of whether and how students differ in their essay composition processes in the presence of a scaffolded, scenario-based assessment design. We examined which writing process features were more important in predicting students' essay performance in a timed writing task between the scaffolded condition and an alternative, reversed condition in which the preparatory tasks followed essay composition. The scaffolded condition was intended to provide the students with a theoretically-grounded assessment design that gave the students ample opportunity to become familiar with the issue at hand and with key aspects of argumentation. The other condition asked the students to write an essay without such guidance and preparation.

The results revealed clear task-ordering effects. Compared to the Reversed Form, the Original Form reduced the dependency of performance on writing fluency. By reducing the importance of general writing fluency, the Original Form appeared to allow students to direct more cognitive resources to editing and revising activities, as well as to the higher-level cognitively demanding tasks important to the quality of the argumentation. In contrast, for students taking the Reversed Form, essay performance was more related to general writing fluency, which would place a burden on less fluent students who would have fewer free cognitive resources for such higher-level writing processes as editing and revision. Students taking this Form also appeared to spend more time immediately before beginning writing, suggesting the need to organize and plan in ways that the Original Form students did not need to do.

This study has several limitations. One limitation is that we used a small convenience sample collected from a single grade level. Hence the findings may not generalize to student populations different from the current one. A second limitation is that because the data were collected under low-stakes testing conditions, students' motivation level might have negatively affected their essay performance and the

effort they expended on writing. A third limitation is that we only analyzed the keystroke features as reported in Zhang & Deane (2015). The inclusion of additional (or different) keystroke features might yield different results, which has particular implications for our findings with respect to the relative importance of the keystroke dimensions. The importance of a dimension was evaluated by the Pratt's measure aggregated across the features that load on that dimension. Pratt's statistic is, in essence, a relative measure, and is sensitive to the predictors entered and their interactions in the regression model. When evaluating our results, criticisms of Pratt's statistic (as well as of other statistics that rely on the variance contribution a predictor makes in a linear model) should be taken into account (see Carlson, 2014 for a comprehensive review of such statistics). A fourth limitation is that the observed between-form differences may not be attributable to scaffolding per se because we did not have a condition (in this chapter) to control for the same task order but without scaffolding. That would be something closer to a form (Form 3) in the original experiential design where the lead-in tasks are taken from a parallel test form on a different topic, or to a form that used one set of sources but lead-in tasks were unrelated to the sources or to the subsequent essay. A final limitation is that we used scores from a single essay as the dependent variable, which is not the most reliable criterion for evaluating writing performance.

As for future research directions, we are at the early stages of understanding how to make meaning from students' writing processes, and how they can be best used in the assessment context. Studies that evaluate the meaning of different features and feature aggregations are needed. Another avenue of future study is to examine the writing processes in the other two assessment forms in the experiment (i.e., one breaking the scenario and the other breaking both the scenario and task ordering, consult Zhang, van Rijn et al., 2016 for details), and compare to the Original Form. Finally, it might be worthwhile to investigate whether the original test design has uniform or differential effects on test takers of different ability levels or demographic backgrounds.

**Acknowledgements** We thank James Carlson, Kathy Sheehan, and Jiangang Hao for their technical review of previous versions of this manuscript. We also thank Randy Bennett for his helpful input and Shelby Haberman and Dan McCaffrey for providing advice on the analyses.

## Appendices

### ***Appendix A: CBAL™ Generic Scoring Guide: Discourse-Level Features in a Multi-paragraph Text (Rubric 1)***

#### **EXEMPLARY (5)**

An EXEMPLARY response meets all of the requirements for a score of 4 but *distinguishes itself by skillful use of language, precise expression of ideas, effective sentence structure, and/or effective organization*, which work together to control the flow of ideas and enhance the reader's ease of comprehension.

#### **CLEARLY COMPETENT (4)**

A CLEARLY COMPETENT response typically displays the following characteristics:

- It is adequately structured.
  - *Overall, the response is clearly and appropriately organized for the task.*
  - *Clusters of related ideas are grouped appropriately and divided into sections and paragraphs as needed.*
  - *Transitions between groups of ideas are signaled appropriately.*
- It is coherent.
  - *Most new ideas are introduced appropriately.*
  - *The sequence of sentences leads the reader from one idea to the next with few disorienting gaps or shifts in focus.*
  - *Connections within and across sentences are made clear where needed by the use of pronouns, conjunctions, subordination, etc.*
- It is adequately phrased.
  - *Ideas are expressed clearly and concisely.*
  - *Word choice demonstrates command of an adequate range of vocabulary.*
  - *Sentences are varied appropriately in length and structure to control focus and emphasis.*
- It displays adequate control of Standard Written English
  - *Grammar and usage follow SWE conventions, but there may be minor errors.*
  - *Spelling, punctuation, and capitalization follow SWE conventions, but there may be minor errors.*

**DEVELOPING HIGH (3)**

A response in this category displays some competence but differs from Clearly Competent responses in at least one important way, including *limited development; inconsistencies in organization; failure to break paragraphs appropriately; occasional tangents; abrupt transitions; wordiness; occasionally unclear phrasing; little sentence variety; frequent and distracting errors in Standard Written English; or relies noticeably on language from the source material.*

**DEVELOPING LOW (2)**

A response in this category differs from Developing High responses because it displays serious problems such as *marked underdevelopment; disjointed, list-like organization; paragraphs that proceed in an additive way without a clear overall focus; frequent lapses in cross-sentence coherence; unclear phrasing; excessively simple and repetitive sentence patterns; inaccurate word choices; errors in Standard Written English that often interfere with meaning; or relies substantially on language from the source material.*

**MINIMAL (1)**

A response in this category differs from Developing Low responses because of serious failures such as *extreme brevity; a fundamental lack of organization; confusing and often incoherent phrasing; little control of Standard Written English; or can barely develop or express ideas without relying on the source material.*

**NO CREDIT (0)**

*Not enough of the student's own writing for surface-level features to be judged; not written in English; completely off topic; or random keystrokes.*

**OMIT**

*Blank.*

***Copyright by Educational Testing Service, 2013 All rights reserved.***

## ***Appendix B: CBAL™ Generic Scoring Guide: Constructing an Argument (Rubric 2)***

### **EXEMPLARY (5)**

An EXEMPLARY response meets all of the requirements for a score of 4 and distinguishes itself with such qualities as insightful analysis (recognizing the limits of an argument, identifying possible assumptions and implications of a particular position); intelligent use of claims and evidence to develop a strong argument (including particularly well-chosen examples or a careful rebuttal of opposing points of view); or skillful use of rhetorical devices, phrasing, voice and tone to engage the reader and thus make the argument more persuasive or compelling.

### **CLEARLY COMPETENT (4)**

The response demonstrates a competent grasp of argument construction and the rhetorical demands of the task, by displaying all or most of the following characteristics:

#### **Command of Argument Structure**

- *States a clear position on the issue*
- *Uses claims and evidence to build a case in support of that position*
- *May also consider and address obvious counterarguments*

#### **Quality and Development of Argument**

- *Makes reasonable claims about the issue*
- *Supports claims by citing and explaining relevant reasons and/or examples*
- *Is generally accurate in its use of evidence*
- *Expresses ideas mainly in the writer's own words*

#### **Awareness of Audience**

- *Focuses primarily on content that is appropriate for the target audience*
- *Expresses ideas in a tone that is appropriate for the audience and purpose for writing*

### **DEVELOPING HIGH (3)**

While a DEVELOPING HIGH response displays some competence, it typically has at least one of the following weaknesses: a vague claim; somewhat unclear, limited, or inaccurate use of evidence; failure to take account of the alternative; noticeable reliance on source language; simplistic reasoning; or occasionally inappropriate content or tone for the audience.



## DEVELOPING LOW (2)

A DEVELOPING LOW response displays problems that seriously undermine the writer's argument, such as a confusing or inconsistent claim; a seriously underdeveloped or unfocused argument; irrelevant confusing, or seriously misused evidence; substantial reliance on source language; an emphasis on opinions or unsupported generalizations rather than reasons and example; or inappropriate content or tone throughout much of the response.

## MINIMAL (1)

A MINIMAL response displays little or no ability to construct an argument. For example, there may be no claim, no relevant reasons and examples, no development of an argument, little logical coherence throughout the response, or mainly use of source language.

## OFF-TOPIC (0)

Consists entirely of source language, is completely off topic, or consists of random key strokes.

***Copyright by Educational Testing Service, 2014 All rights reserved.***

## References

- Bennett, R. E., Deane, P., & van Rijn, P. W. (2016). From cognitive domain theory to assessment practice. *Educational Psychologist, 51*, 82–107.
- Bruner, J. S. (1978). The role of dialogue in language acquisition. In A. Sinclair, R. J. Jarvella, & W. J. M. Levelt (Eds.), *The child's concept of language*. New York, NY: Springer.
- Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin, 114*, 542–551.
- Carlson, J. E. (2014). *A generalization of Pythagoras's theorem and application to explanations of variance contributions in linear models* (Research Report 14–18). Princeton, NJ: Educational Testing Service.
- Chase, C. I. (1960). Computation of variance accounted for in multiple correlation. *The Journal of Experimental Education, 28*, 265–266.
- Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks* (Research Report 14–03). Princeton, NJ: Educational Testing Service.
- Deane, P., Sabatini, J. P., Feng, G., Sparks, J., Song, Y., Fowles, M., O'Reilly, T., Jueds, K., Krovetz, R., & Foley, C. (2015). *Key practices in the English language arts (ELA): linking learning theory, assessment, and instruction* (Research Report 15–17). Princeton, NJ: Educational Testing Service.

- Deane, P., & Zhang, M. (2015). *Exploring the feasibility of using writing process features to assess text production skills* (Research Report 15–26). Princeton, NJ: Educational Testing Service.
- Greenfield, P. M., & Smith, J. H. (1976). *Structure of communication in early language development*. New York, NY: Academic Press.
- Guo, H., Deane, P., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2016). *Exploring the heavy tailed key-stroke data in writing processes*. Manuscript submitted for publication.
- Jackson, S. L., Krajcik, J., & Soloway, E. (1998). *The design of guided learner-adaptable scaffolding in interactive learning environments*. In *Proceeding of the SIGCHI conference of human factors in computing systems* (pp. 187–194). New York, NY: ACM Press/Addison-Wesley Publishing Co.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY/UK: Macmillan Publishing Co Inc..
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
- Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences*, *13*, 423–451.
- Puntambekar, S., & Hubscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist*, *40*, 1–12.
- Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In T. Pukilla & S. Duntaneu (Eds.), *Proceedings of second tampere conference in statistics* (pp. 245–260). Tampere, Finland: University of Tampere.
- Sabatini, J. P., O'Reilly, T., Halderman, L., & Bruce, K. (2015). Broadening the scope of reading comprehension using scenario-based assessments: Preliminary findings and challenges. *L'Année Psychologique*, *114*, 693–723.
- Sheehan, K. S., & O'Reilly, T. (2012). The case for scenario-based assessments of reading competency. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Reaching an understanding: Innovation in how we view reading assessment*. Lanham, MD: R&L Education.
- Shepard, L. A. (2005). Linking formative assessment to scaffolding. *Educational Leadership*, *63*(3), 66–71.
- Stone, E. (2016). *Integrating digital assessment meta-data for psychometric and validity analysis*. Manuscript submitted for publication.
- Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, *45*, 253–275.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.). Cambridge, MA: Harvard University Press.
- Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership*, *49*(8), 26–33.
- Zhang, M. & Deane, P. (2015) *Process features in writing: internal structure and incremental value over product features* (Research Report 15–27). Princeton, NJ: Educational Testing Service.
- Zhang, M., van Rijn, P. W., Deane, P., & Bennett, R. E. (2016). *Scenario-based assessments in writing: An experimental study*. Manuscript submitted for publication.
- Zhang, M., Hao, J., Li, C., & Deane, P. (2016). Classification of writing patterns using keystroke logs. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (pp. 299–314). New York, NY: Springer.

# Chapter 18

## National and International Educational Achievement Testing: A Case of Multi-level Validation Framed by the Ecological Model of Item Responding

**Bruno D. Zumbo, Yan Liu, Amery D. Wu, Barry Forer,  
and Benjamin R. Shear**

The results of large-scale student assessments are increasingly being used to rank nations, states, and schools and to inform policy decisions. Traditionally, educational testing and assessment in the domains of science and mathematics have typically focused on assessment of learning (i.e., summative) or even assessment for learning (i.e., formative), with inferences regarding a student's individual learning or knowledge as the focus. National and international educational testing programs such as the Trends in International Mathematics and Science Study (*TIMSS*) and the National Assessment of Educational Progress (*NAEP*), on the other hand, are designed to inform policy and assess the impact of community-scale interventions and changes by making inferences about groups of students. As such, although *TIMSS* and *NAEP* measure individual student learning and knowledge, these assessments are neither designed to, nor able to provide, any feedback about individual students. *TIMSS* and *NAEP* are thus quintessential candidates for what Zumbo and Forer (2011) and Forer and Zumbo (2011) refer to as multilevel assessments of multilevel constructs.

---

B.D. Zumbo (✉) • Y. Liu • A.D. Wu  
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: [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca); [yan.liu@ubc.ca](mailto:yan.liu@ubc.ca); [amery.wu@ubc.ca](mailto:amery.wu@ubc.ca)

B. Forer  
The Human Early Learning Partnership, The University of British Columbia,  
Suite 440, 2206 East Mall, Vancouver, BC V6T 1Z3, Canada  
e-mail: [barry.forer@ubc.ca](mailto:barry.forer@ubc.ca)

B.R. Shear  
School of Education, University of Colorado Boulder, 249 UCB, Boulder, CO 80309, USA  
e-mail: [benjamin.shear@colorado.edu](mailto:benjamin.shear@colorado.edu)

As Zumbo and Forer (2011) note, a multilevel construct can be defined as a phenomenon that is potentially meaningful both at the level of individuals and at one or more levels of aggregation, but the construct is interpreted and used primarily at the aggregate level. While all constructs reside at one level at least, an organizational setting like formal education is inherently multilevel, given the natural nesting of students within classes within schools within school districts. Having to deal with multilevel issues should be assumed when studying phenomena in these multilevel settings (e.g., Klein, Dansereau, & Hall, 1994; Morgeson & Hofmann, 1999).

Building on Zumbo's (2007, 2009) view of validity as contextualized and Pragmatic explanation, we will demonstrate the validation process of multilevel educational achievement constructs using the 2007 *TIMSS* mathematics assessment – although we do not do so herein, similar methods could be used with other assessments, for example *NAEP*. Multilevel validation methods aim to provide a strong form of construct validity; that is, the evidence should provide an explanation for the observed variation in test scores. From this explanation-focused point of view, validity involves inference and the weighing of evidence, guided by explanatory goals. According to Zumbo's view, explanation acts as a regulative ideal; validity is the explanation for the test score variation, and validation is the process of developing and testing the explanation.

In the context of multilevel assessments and constructs, the multilevel nature of the assessment (and of the data) must guide the explanatory model. Therefore, a multilevel explanation which considers both the complex structure and nature of the data, as well as the contextual factors at the various levels of aggregation, is a defining feature of multilevel validity and multilevel validation that support the inferences we make from these scores. We frame multilevel validity within an ecological model of item responding (Zumbo et al., 2015). The ecological model of item responding takes a Pragmatic stance and begins with the statement that an explicit recognition of the complexity of our world and the need to be sensitive to context necessitates research methods for validation that are sturdy and robust to the idiosyncrasies of real world data that are often complex and involve multiple levels of variables – personal, school, community, and regional variables all acting simultaneously. As an explanatory model of test score variation, our multilevel validation framework is embedded within an ecological model of item responding that is situated within a Pragmatic view of abductive explanation wherein one develops validity evidence for tests through abductive reasoning (Stone & Zumbo, 2016; Zumbo, 2007, 2009). In contrast to inductive reasoning or deductive reasoning, abductive reasoning neither construes the meaning of the scores purely from empirical evidence nor assumes the meaning of the test in order to explain the score. Rather, abductive reasoning seeks the enabling conditions under which the score makes sense. This serves as the conceptual foundation for our approach to multilevel test validation.

Zumbo and colleagues (Zumbo & Gelin, 2005; Zumbo et al., 2015) have suggested that to understand the item responses, different explanatory sources, such as

psychological and cognitive factors, physical and structural settings of the community, as well as the social context need to be explored. Viewed in an ecological framework, item responses and test performance cannot be simply attributed to the individuals or the environment, but to the relationship between the two.

We approach the validation process using Chen et al.'s (2004a, 2004b) adapted step-by-step procedures for conducting multilevel construct validation. More specifically, we use two complementary explanatory methods in the validation process: multilevel latent variable modeling techniques (e.g., Kaplan & Elliott, 1997; Muthén, 1994) and within-and-between analysis (WABA; Dansereau & Yammarino, 2000).

*TIMSS* data are commonly used to rank nations based upon their aggregate scores; therefore, validation of these scores must be appropriate to this aggregate level of data. From Zumbo's explanation focused view of validity, the validity of the inferences one can make from the multilevel assessment data depends on explaining the variation in the aggregate level data. Our aim is to provide empirical support for the inferences made from the multilevel test scores by developing and testing empirical models that can account for the national level variation in mathematics achievement. The explanation focused view of validity accompanied by the ecological model of item responding (Zumbo et al., 2015) situates conventional response process research in a multilevel construct setting and moves response process studies beyond the traditional focus on individual test-takers' behaviors.

## Method

### *Measure and Sample*

According to Mullis et al. (2005), *TIMSS* 2007 was the fourth in a continuing cycle of curriculum-based international assessments in mathematics and science. Mullis et al. go on to state that the target population of *TIMSS* 2007 was all students at the end of Grades 4 and 8 in the participating countries. According to the *TIMSS* assessment framework, the grade 8 sample includes children aged 13 and 14, and is defined as the upper of the two adjacent grades with the most 13-year-olds.

Our analyses focused on grade eight 2007 *TIMSS* mathematics achievement. The analyses were conducted on booklet one. The eighth grade mathematics content domains are:

- (a) number with 11 items,
- (b) algebra with 8 items,
- (c) geometry with 6 items, and
- (d) data and chance with 4 items.

Four of the 29 items were polytomous (scored on a 0, 1, 2 scale) and the remainder were binary (scored as 0 or 1). There were two polytomous items in the number domain, one in the geometry domain, and one in the data and chance domain. The four domain scores, recorded as percentages, were computed and used as continuous indicators of mathematics achievement in the statistical analyses. The analyses involved a total of 15,529 students, from 48 nations ranging from 234 to 544 students per nation – the average number of students per nation was 323.5. A list of all the countries involved in this study, and their corresponding number of students in this study, can be found in Appendix A.

### *National Level Explanatory Variables*

As a reminder, from our explanation focused view of validity, explaining the variation in the aggregate level data goes a long way toward establishing the validity of the inferences one can make from the multilevel assessment data. In addition to assessing mathematics and science achievement, the *TIMSS* program of studies collects background data from the *TIMSS* research coordinator within each nation. In particular, we were interested in *TIMSS*'s Curriculum Questionnaire. The Curriculum Questionnaire is primarily centered on the defined national or regional curriculum in eighth grade, including what it prescribed and how it is disseminated. Appendix B contains a list of the national level explanatory variables as well as the variable description and data coding. We used this national data provided by *TIMSS* with an eye toward explaining the variation in the aggregate level mathematics achievement data.

We wanted our explanatory models to reflect that education does not exist in a vacuum and may be influenced by (or reflect) various national socio-economic differences. Therefore, in addition to the national curriculum variables we also included national level explanatory variables characterizing national social conditions and processes. *TIMSS* reports on several such national social indicators, including life expectancy and the Human Development Index (HDI) for each nation. Life expectancy reflects the health and wellbeing of the nation and is a standard social indicator. The HDI is based on Mahbub ul Haq and Amartya Sen's human development approach and, in short, measures the average achievements in a nation on three basic dimensions of human development: a long and healthy life, knowledge, and a decent standard of living. As described in technical notes of the Human Development Report (Watkins, 2007), health is measured by life expectancy at birth; knowledge is measured by a combination of the adult literacy rate and the combined primary, secondary, and tertiary gross enrolment ratio; and standard of living is measured by GDP per capita (PPP US\$).

## ***Validation Method and Data Analyses***

As described in Zumbo and Forer (2011) we followed Chen et al.'s (2004a, 2004b) adapted step-by-step procedures for conducting multilevel construct validation. Given that this is an expository essay of a validation method, rather than a study of *TIMSS*, per se, the results section will provide further details about the validation methods and results. In short, however, Chen et al.'s approach generally involves four steps. The first step focuses on the construct definition at each level and the nature of the construct at the aggregate level(s). The second step requires that one be explicit about the nature and structure of the aggregate construct; that is, one needs to select an appropriate composition model. The third step focuses on the psychometric properties across levels and usually involves multilevel modeling – in our case multilevel factor analysis and/or item response modeling. In addition, as part of this third step, multilevel conditional models are explored to investigate whether national level variables contribute to the model. The fourth step has one focus on the construct variability within and between units and has on ensure that there is sufficient variability within and between units (i.e., at lower and higher levels). The sorts of statistics used at this step allow one to investigate whether, for some aggregate-level measures, inter-member reliability (intra-class correlations, ICCs) can provide relevant evidence. In addition a within-and-between *multiple relationship* analysis is reported (complementary to the conditional models of the multilevel analysis) to investigate whether national level variables moderate the findings.

## **Results and Conclusions**

The findings will be organized according to the Chen four-step framework.

### ***Steps 1 and 2: Construct Definition and Aggregate Model***

The first two steps are conceptual rather than data-driven, dealing with (i) the theoretical issues of construct definition (such as the construct's domain postulated dimensionality), and (ii) describing the nature of the aggregate construct. In terms of construct definition in step one, *TIMSS* is designed to align broadly with mathematics curricula in the participating countries. Therefore, at the student and national levels the results are meant to reflect the degree to which students have learned mathematics concepts and skills likely to have been taught in school. From a multilevel validity point of view, however, the question to be addressed in steps three and four (below) are whether the national differences also reflect other secondary national attributes; for example, curricular differences and/or socio-economic differences which may reflect differences in opportunities to learn. In terms of step



two, *TIMSS* uses an aggregate, global summary index model (the national average) that describes the group (nation) as a whole, and hence summarizes the collection of lower level, individual student mathematics scores (Hofmann & Jones, 2004).

### ***Step 3: Psychometric Properties Across Levels***

The psychometric analyses across levels were conducted in three phases: exploratory multilevel item level analyses for each domain, confirmatory multilevel factor analyses without any predictors at the within or between levels, and finally, confirmatory multilevel factor analyses with national level predictors at the between level. A compact general model for characterizing multilevel factor analyses (of the exploratory or confirmatory variety) in our case is:

$$\Sigma_T = \Sigma_B + \Sigma_W, \quad (18.1)$$

wherein, the between-nations factor model is hypothesized to account for the covariance structure associated the between-nations random components  $\Sigma_B$ , and the within-nations factor model is hypothesized to account for the covariance structure associated with the within-nation random components  $\Sigma_W$  (e.g., Goldstein & McDonald, 1988; Kaplan & Elliot, 1997; Lee, 1990; Longford & Muthén, 1992; Muthén, 1994; Muthén & Satorra, 1995). Following Muthén (1994) a general form of the multilevel factor analysis model can also be written as,

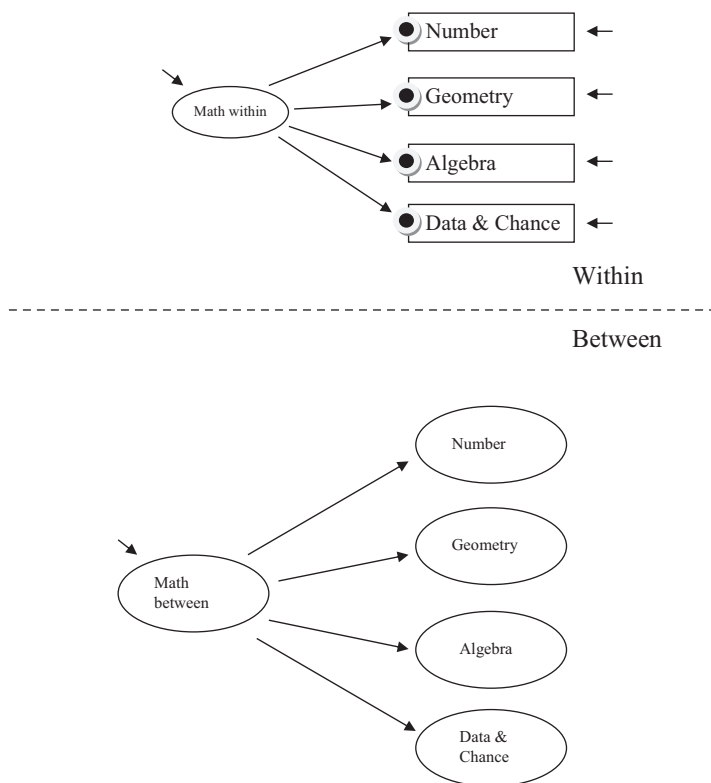
$$X_{ij} = \alpha + \Lambda_B \xi_{Bj} + \Lambda_W \xi_{Wij} + \delta_{Bj} + \delta_{Wij}, \quad (18.2)$$

where  $X_{ij}$  is the vector of indicator scores for the  $i$ th student in the  $j$ th nation,  $\alpha$  is a vector of observed variable measurement intercepts,  $\Lambda_B$  and  $\Lambda_W$  denote the loading matrices for the between and within factors,  $\xi_{Bj}$  and  $\xi_{Wij}$ , respectively. Finally,  $\delta_{Bj}$  and  $\delta_{Wij}$  denote the between-nation and within-nation error variables. The concepts in Eqs. (18.1) and (18.2) are the driving engine behind the exploratory and confirmatory factor models used herein.

Given that we were using domain scores in our subsequent analyses, in the first phase we confirmed that each domain was unidimensional. The dimensionality of each domain was investigated using a multilevel exploratory factor analysis at the item level – in essence, a multilevel exploratory item response theory analysis. The possibility of one or more factors within and between (allowing for different number of factors across levels) was investigated using *Mplus* 5.2 software. Table 18.1 presents the eigenvalues for the within and between correlation matrices, as well as the fit statistics for the one factor within and one factor between model – both the eigenvalues greater than one and the ratio of the first to second eigenvalue greater than three were investigated; as well as whether the CFI was greater than 0.95, RMSEA was less than 0.08, and the SRMR (within and between) were less than

**Table 18.1** Eigenvalues and fit indices of multilevel item exploratory factor analyses for each domain

Eigenvalue	Number		Geometry		Algebra		Data & chance	
	Within	Between	Within	Between	Within	Between	Within	Between
1st	3.89	8.72	3.29	6.26	3.21	5.01	2.17	3.43
2nd	1.04	0.66	0.93	0.52	1.04	0.54	0.91	0.37
3rd	0.90	.040	0.85	0.47	0.72	0.21	0.54	0.17
SRMR	0.03	0.05	0.03	0.05	0.09	0.05	0.05	0.01
CFI	0.99		0.99		0.97		0.97	
RMSEA	0.01		0.01		0.07		0.04	



**Fig. 18.1** Multilevel confirmatory factor analysis of the four continuous domains

0.10. In all cases, a one factor model within and between was supported and therefore each domain score could be used in subsequent analyses.<sup>1</sup>

<sup>1</sup>Domain scores were used in the multilevel confirmatory factor analyses because they could be treated as continuous observed variables and hence conventional fit statistics were available to test for fit as well as the computational ease of using continuous scores resulting in substantially reduced computing time. Our's is a variation on the use of item parcels. In our case, however, the

In the second phase, a two-level confirmatory factor model of the four continuous indicators was fit using *Mplus 5.2* with student responses as level one, and nation/state as level two – using maximum likelihood estimation with robust standard errors (MLR). Figure 18.1 depicts the multilevel confirmatory factor model in path diagrammatic notation. The upper part of the diagram depicts the within (student) level and the hypothesized factor (math). It is important to note that the filled circles at the observed domain indicate the random measurement intercepts.

On the between level these random intercepts are continuous latent variables varying over nations, where the between-nation variation and covariation are represented by the nation-level math factor. The meaning of the student-level factor is akin to that with regular factor analysis. In contrast, the between-level math factor represents the national-level phenomena for which a researcher typically has less understanding and hence is the focus of the explanatory focus in our multilevel validation framework.

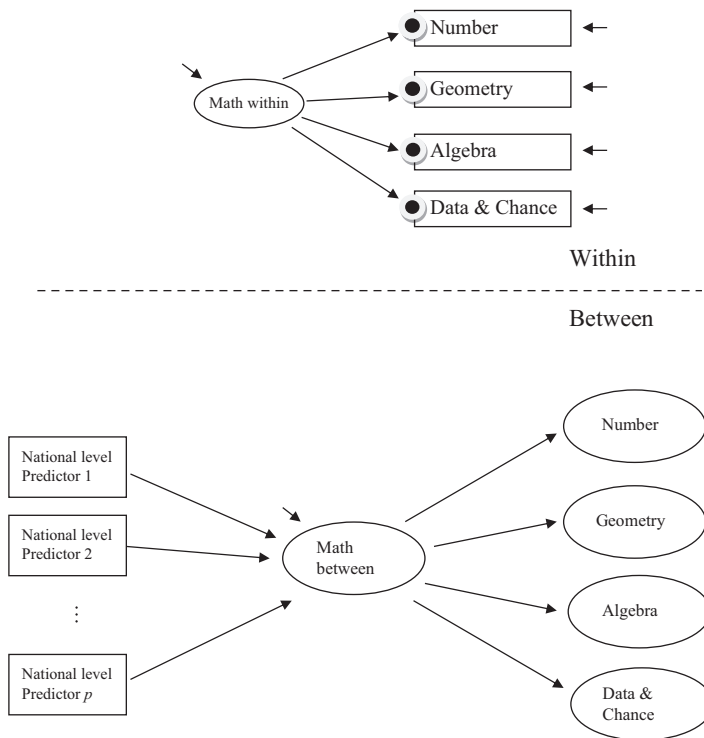
In Steps 1 and 2 of the validation framework we articulated that this latent variable is often considered to be “mathematics achievement” and is reported as an aggregate global summary index (the national average score on mathematics achievement). This national level latent variable, used to rank nations, is the focus of much of the *TIMSS* reporting and policy research. As noted above, from a multi-level validity point of view, questions arise as to whether the national mathematics score differences reflect not only “individual student mathematics achievement,” but other secondary national attributes such as curricular differences and/or socioeconomic differences, which may reflect differences in opportunities to learn.

The two-level model depicted in Fig. 18.1 showed good fit: CFI = 1.0, TLI = 0.999, RMSEA = 0.01, and the SRMR within of 0.002, and between of 0.008. The loadings were all statistically significant and the standardized coefficients ranged from 0.658 to 0.808 (within), and 0.936 to 0.983 (between). The estimated intra-class correlations (ICC) for the four mathematics domains were: number ICC = 0.344, geometry ICC = 0.282, algebra ICC = 0.260, and data and chance ICC = 0.286. Together this evidence supports the need to take into account the level two (nation) in modeling the assessment results.

Given the results of the above (unconditional) two-level model, national level predictors were added. Figure 18.2 depicts the two-level factor analysis model with national predictors at level two (between nations). As a reminder, there were two

---

parcels are theoretically driven and confirmed to be unidimensional. As further support for the use of the four domain scores in subsequent analyses, we fit a multilevel exploratory item response theory analysis for all 29 items simultaneously. The first three eigenvalues of the within level polychoric correlation matrix were 10.0, 1.5, and 1.3; and the first three eigenvalues of the between level correlation matrix were 22.4, 1.5, and 1.0. Clearly, the eigenvalues point toward one between and one within latent variable even when the items are the focus of analysis. The CFI = 0.92, RMSEA = 0.03, SRMR Within = 0.07, and SRMR Between = 0.06 for the one factor within and one factor between model. As an example of the computational burden of the item level analyses, the 29 item analysis described in this footnote required over 6 h of computational time whereas the domain models complete in less than 5 min each. All of this evidence lends further support for the use of the domain scores in the subsequent analyses.



**Fig. 18.2** Multilevel confirmatory factor analysis of the four continuous domains with national level predictors

categories of national level predictors: (a) the national curriculum questions in Appendix B, and (b) the national life expectancy and HDI social indicators.

The strategy for testing the significance of the potential predictors, with an eye toward selecting a final model, began by running all the single-predictor models in the multilevel factor analysis – i.e., the simple regressions of the between nation latent variable on each of the candidate variables listed in Appendix B, as well as life expectancy and HDI, individually. From these simple models we selected a set of predictors that were then submitted to a multiple predictor multilevel factor analysis model. Given that the HDI has an “educational” component, at this stage the multiple predictor models were fit separately for the curriculum and social indicator variables – the concern being that the educational component of the HDI would “swamp” the effect of the curriculum variables. In the third stage, only those predictors that were significant in the multiple predictor models were chosen for the final model. With the final model at hand (assuming that this final model included more than one predictor) latent variable Pratt indices (Zumbo, 2007) were computed to order the national level predictor variables in terms of the proportion of the between level latent variable model R-squared that each accounted for.

<b><i>Statistically Significant Predictors of the Between-nation mathematics latent variable</i></b>
Calculator
Computer
Communicate
Reason
Integrate (negative sign)
Life expectancy
HDI
<b><i>NOT Statistically Significant Predictors of the Between-nation mathematics latent variable</i></b>
Basic
Concept
Real life
Proof
Degree
Remedial
Exam
DfCur and DfLevel... two variables which describe how the mathematics curriculum addresses the issue of students with different levels of ability.

**Fig. 18.3** Summary listing of the single predictor models

Figure 18.3 provides a summary of the single predictor models based on the multilevel factor model in Fig. 18.2. There are several noteworthy findings among the single predictor models. It is important to note that both the statistically significant and non-significant explanatory variables are relevant to the eventual validation conclusions. Statistically non-significant explanatory variables need to be reported in validation studies so as to avoid the deleterious effects of the so-called “file drawer” problem. In short, non-significant explanatory variables can help shape the validity conclusions. For example, we found that the variation in the nation-level mathematics achievement measure in TIMSS *does not reflect*: (a) whether there are national administered examinations in mathematics that have consequences for individual students, such as determining grade promotion, entry to a higher school system, entry to a university, and/or exiting or graduating from high school, or (b) whether one needs a degree from a teacher education program to be a middle/lower secondary grade teacher. These are noteworthy findings because neither of these between-nation variables was statistically significant predictors of the between-nation variation in mathematics scores although both have been discussed in the assessment literature as potential confounding variables of *TIMSS* rankings of nations. Likewise, a surprising finding is that there is a statistically significant negative coefficient for the ‘integrate’ variable suggesting that a national policy emphasizing the integration of mathematics with other subjects results in a lower between national mathematics score.

Next, a model was fit with the statistically significant curriculum variables from the simple predictor models in Fig. 18.3. A model with multiple national level predictors (Fig. 18.2) was fit. Table 18.2 shows the fit statistics for this model and the significance tests for each of the national level predictors. We can see from Table 18.2 that the calculator and reason variables were statistically non-significant in the

**Table 18.2** Model results of the curriculum national level predictors

MATH (between) ON	Estimate	S.E.	Est./S.E.	Two-tailed P-Value	Significant?
Calculator	-3.250	3.108	-1.046	0.296	n.s.
Computer	12.462	3.642	5.422	0.001	Significant
Communicate	5.664	2.713	2.088	0.037	Significant
Reason	0.990	2.272	0.436	0.663	n.s.
Integrate	-9.761	2.257	-4.324	0.000	Significant

Note: The model R-square is 0.438, which is statistically significant,  $z = 6.129$ ,  $p < .0001$ . Model fit: RMSEA = 0.009, SRMR within = 0.001 and SRMR between = 0.028

**Table 18.3** Model results of the social indicator predictors

MATH (between) ON	Estimate	S.E.	Est./S.E.	Two-tailed P-Value	Significant?
Life expectancy	-0.311	0.330	-0.942	0.346	n.s.
HDI	111.478	21.287	5.237	0.000	Significant

Note: The model R-square is 0.431, which is statistically significant,  $z = 5.419$ ,  $p < .0001$ . Model fit: RMSEA = 0.013, SRMR within = 0.001 and SRMR between = 0.029

**Table 18.4** Results of the final model, including the Pratt Index

MATH (between) ON	Standardized estimate	S.E.	Est./S.E.	Two-tailed P-value	Simple regression estimate	Pratt Index
Computer	0.236	0.092	2.557	0.011	0.429	0.19
Integrate	-0.285	0.101	-2.818	0.005	-0.370	0.20
HDI	0.498	0.080	6.231	0.000	0.653	0.61

Note: The model R-square is 0.533, which is statistically significant,  $z = 6.461$ ,  $p < .0001$ . Model fit: RMSEA = 0.009, SRMR within = 0.001 and SRMR between = 0.024

presence of the other curriculum variables. Likewise, Table 18.3 shows the fit statistics and significance tests for the two social indicators; with only the HDI having been statistically significant.

With the results of Tables 18.2 and 18.3 at hand, the model in Fig. 18.2 was fit with computer, communicate, integrate, and the HDI as national level predictors. That model was found to fit the data but that communicate was statistically non-significant. Therefore, the final model included: computer, integrate, and the HDI (Table 18.4).

The final model tells us that 53.3% of the between-nation variation constructed from the measurement model can be accounted for by:

- (i) whether the national curriculum contains statements/policies about the use of computers in grade 8 mathematics,
- (ii) how much emphasis the national mathematics curriculum places on integrating mathematics with other subjects, and
- (iii) the Human Development Index, HDI.

Interestingly, 61% of the model R-squared can be attributed to the HDI making it the most important predictor. Curiously, the negative regression coefficient for integrating mathematics with other subjects seems to suggest that the more one integrates mathematics with other subjects the lower the between-nation latent variable score. A possible explanation for this finding is that there may be a trade-off being made when one integrates the mathematics so that the students get less exposure to basic mathematics (this is, of course, speculation). Additionally, the questionnaire does not distinguish between integrating mathematics into other courses versus integrating other subjects into the mathematics classroom, which impact student learning differently. One should, of course, be cautious in interpreting these findings because they are correlational (the ‘predictors’ are, more formally, ‘covariates’), and therefore a third variable(s) may be the source of the covariation.

#### ***Step 4: Construct Variability Within and Between Units***

The fourth step in the multilevel construct validation process is an analysis of the relative amounts of within-group and between-group variation, which provides empirical guidance about appropriate levels of aggregation. In short, it addresses the question of how valid it is to mathematics achievement in terms of averaged student test score performance. We addressed this question with two analytic methods: mixed effects modeling and within-and-between analysis.

**Mixed Effects Modeling: Reliability and Multilevel Measurement** Raudenbush, Rowan, and Kang (1991) discussed several key issues involved in multilevel measurement. As they show, one convenient way to model multilevel measurement data is to fit a three-level multilevel regression model with separate levels for the domains, students, and nations. Using a model with no explanatory variables except the intercept one has, using Raudenbush’s notation:

$$\begin{aligned}
 \text{Level One : } y_{ijk} &= \pi_{0jk} + e_{ijk}, & e_{ijk} &\sim N\left[0, \text{var}(e_{ijk})\right] \\
 \text{Level Two : } \pi_{0jk} &= \beta_{00k} + r_{0jk}, & r &\sim N\left[0, \text{var}(r_{0jk})\right] \\
 \text{Level Three : } \beta_{00k} &= \gamma_{000} + u_{00k}, & u_{00k} &\sim N\left[0, \text{var}(u_{00k})\right] \\
 \text{Combined Equation : } y_{ijk} &= \gamma_{000} + r_{0jk} + u_{00k} + e_{ijk}, & & (18.3)
 \end{aligned}$$

wherein  $\gamma_{000}$  is the intercept term,  $i$  denotes domains,  $j$  students, and  $k$  nations. The three level model in Eq. (18.3) was fit using HLM version 6 and variance components for domains, students, and nations were 320.46, 279.32, and 214.27, respectively. In short, the domain-variance component due to domain score inconsistency,



student-variance component is an estimate of the variation of the mean domain score between students within the same nation, the nation-variance component is an estimate of the variation of the mean domain score between different nations. The intercept term, the grand mean, was 41.0 which indicates that the overall average mathematics score was 41.0%. The error variance in the mean of the four mathematics domains was 80.12 (error standard deviation was 8.95). In addition, the student-level internal consistency was 0.78 and the nation-level internal consistency was 0.99. The student-level internal consistency shows that the student-level variability is not random error, but that it is systematic. Likewise, it should be noted that the nation-level internal consistency depends on (i) the number of domains in the test, (ii) the average correlation among the domains at the national level, (iii) the number of students sampled in the nation, and (iv) the intraclass correlation at the national level.

Please note that the domain-level portion of Eq. (18.3), level one, is included in the model only to produce an estimate of the domain-variance component. Equation (18.3) is akin to a multilevel factor analysis, as described in Eqs. (18.1) and (18.2) and depicted in Fig. 18.2, except that the factor loadings would be constrained to be equal across the four domains both for the within and between latent variables. Clearly, Eq. (18.3) is a limited model compared to the multilevel factor analysis.

**Within-and-Between Analyses** Within-and-between analysis (WABA; Dansereau & Yammarino, 2000) is an alternative multilevel validation technique that compares patterns of within- and between-group variability to determine appropriate levels of aggregation, using tests of both statistical and practical significance. The mathematical engine of the WABA is the so-called covariance theorem which states that the student-level covariance between two variables,  $x$  and  $y$ , is sum of the nation-level covariance of  $x$  and  $y$ , and the sum of the  $J$  within-nation covariances of  $x$  and  $y$ . This can be expressed more formally as

$$\text{cov}(x,y) = \text{cov}(x_j,y_j) + \sum_{j=1}^J \text{cov}(x_j,y_j),$$

for the  $j$ th nation, indexed by  $j = 1, 2, \dots, J$ . The fundamental equation of WABA analyses can be written as:

$$r_{xy} = \eta_{Bx}\eta_{By}r_{Bxy} + \eta_{Wx}\eta_{Wy}r_{Wxy},$$

wherein  $r_{xy}$  denotes the total Pearson correlation between  $x$  and  $y$ ;  $r_{Bxy}$  and  $r_{Wxy}$  denote the between group and within groups Pearson correlation of  $x$  and  $y$ , respectively; and  $\eta_{Bx}$ ,  $\eta_{By}$ ,  $\eta_{Wx}$ ,  $\eta_{Wy}$  denote the between-groups and within-groups eta correlation ratio (also called the eta coefficient) for  $x$  and  $y$ , respectively.

Without empirical testing it is difficult to determine if a within-nation correlation or a between-nation correlation is more descriptive of a data set. Thus, we used a within and between analysis (WABA) technique that, unlike multilevel (HLM) analysis, makes no a priori assumption about the level(s) at which inferences are most appropriate. In this respect, WABA provides important guidance for later HLM analyses, which typically are concerned with explaining multilevel effects on indi-

**Table 18.5** Summary of the WABA analyses

Domain	Eta-correlation between	Eta-correlation within	E-ratio
Number	0.589	0.808	0.728
Geometry	0.529	0.849	0.624
Algebra	0.506	0.862	0.587
Data & Chance	0.536	0.844	0.635

**Table 18.6** Within and between-nation domain correlations and correlation decomposition

Domain correlations	Between-nation correlation	Within-nation correlation	Overall individual-level correlation	Between-nation component	Within-nation component
Number-Geometry	0.956	0.648	0.743	0.298	0.445
Number-Algebra	0.910	0.551	0.655	0.271	0.384
Number-Data	0.957	0.528	0.662	0.302	0.360
Geometry-Algebra	0.930	0.541	0.645	0.249	0.396
Geometry-Data	0.938	0.529	0.645	0.266	0.379
Algebra-Data	0.892	0.449	0.569	0.242	0.327

viduals. If individual-level inferences are appropriate according to WABA, this justifies the inherent levels assumptions in HLM, which can then be used to estimate specific effects (something that WABA is not designed to do).

Tables 18.5 and 18.6 summarize the most pertinent results of the WABA analyses. Table 18.5 focuses on the within- and between-nation variance for the four domains, while Table 18.6 focuses on the within- and between-nation covariances for the same domains. For each of the domains, the preponderance of variance is within-nation, as reflected in the low E-ratios. However, for the covariances across each pair of domains, the between-nation correlation is higher than the within-nation correlation. So far, this argues for a weak inference that the most appropriate level is the national level. However, the relatively large within-nation correlations imply that the stronger inference would be that the individual level is most appropriate. This conclusion is also supported by the decomposition of the individual-level correlations, shown in Table 18.6. Across all pairs of domains, both the between- and within-nation components contribute to the overall correlation, which is consistent with an individual-level inference.

The potential moderating effect of each of the three significant nation-level covariates (HDI, computer, and integrate) was also tested using WABA multiple relationship analysis, which tests whether the pattern of within- and between-nation variances and covariances is different across categories of each covariate. If the WABA-inference is different for one or more categories of a covariate, a moderation effect is assumed. Both the Computer and Integrate variables are binary. The

continuous HDI scores were converted to a three-category version (low, medium high) with approximately equal number of individuals per category.

The WABA multiple relationship analyses did not show any moderating effect for any of the three nation-level variables. Across all categories of these variables, the same individual-level inference was found.

## Conclusions

The purpose of this paper was to demonstrate multilevel validation in the context of large-scale international (and national) educational achievement testing such as *TIMSS* or *NAEP*. Focusing on *TIMSS* and using an explanatory-focused framework for multilevel measurement validation (Forer & Zumbo, 2011; Zumbo, 2009; Zumbo & Forer, 2011), our aim was to provide empirical support for the inferences made from the test scores by developing and testing empirical models that account for the national level variation in mathematics achievement. *TIMSS* data are commonly used to rank nations using the national level aggregate score. In this context, the goal of validation is to explain the variation at this aggregate level of data. Multilevel explanatory validation is a novel type of response process validation research.

Using Chen et al.'s (2004a, 2004b) adapted step-by-step procedures for conducting multilevel construct validation, we were able to use multilevel modeling and WABA (in the fourth step) to establish that it is valid to discuss mathematics achievement in terms of averaged student test score performance at the national level; the WABA, however, cautioned us that although meaningful, this between nation aggregation is only marginally justified.

In the third step of Chen et al.'s procedures, we were able to investigate the sources of the between nation variation. In particular, we investigated whether variables external to the testing environment itself, such as curricular and socioeconomic variables, might be useful explanatory sources of the between nation variation (and hence, perhaps, of the international rankings themselves). It is appropriate at this point to caution against making causal inferences about national level predictors. With that limitation in mind, however, one can conclude from our explanatory-focused view of multilevel validity (Zumbo, 2009) that the national level mathematics measure reflects (i) the degree to which students have learned mathematics concepts and skills likely to have been taught in school, (ii) differences between national curricula regarding the use of computers in grade 8 mathematics and the emphasis placed on integrating mathematics with other subjects, and finally, (iii) national variation on social indicators of a long and healthy life, knowledge, and a decent standard of living – as measured by the Human Development Index. Using an explanation-focused view of validity, we found that explaining variation in the aggregate level data goes a long way toward establishing the validity of the inferences one can make from multilevel assessment data.

## Future Directions

There are two very important caveats to our findings. First, the analyses were conducted in an unweighted manner and hence did not reflect the complex sampling plan used by *TIMSS*. Second, the HDI as a composite index that needs to be disaggregated to more fully investigate its relation with *TIMSS* mathematics achievement scores. Our next steps include tackling the complex weighting issue and exploring the disaggregated HDI data.

**Acknowledgment** The authors would like to thank Professor Fred Dansereau for his generous guidance and feedback on the WABA analyses, and Professor Bob Linn for the encouragement to publish this paper. An earlier version of this paper presented at the symposium “A Multilevel View of Test Validity”, 2010 Annual Meeting of the American Educational Research Association, Denver, CO.

## Appendices

### *Appendix A: Countries Involved in the Study and Sample Size*

Nation	Number of students
Algeria	384
Armenia	277
Australia	294
Bahrain	303
Bosnia and Herzegovina	301
Botswana	298
Bulgaria	288
Chinese Taipei	287
Colombia	347
Cyprus	314
Czech Republic	349
Egypt	466
England	299
Georgia	306
Ghana	377
Hong Kong, SAR	249
Hungary	285
Indonesia	305
Iran, Islamic Republic of	291
Israel	234
Italy	315
Japan	307
Jordan	370
Korea, Republic of	306
Kuwait	284
Lebanon	267
Lithuania	287
Malaysia	321
Malta	337
Mongolia	317
Norway	326
Oman	322
Palestinian National Authority	315
Qatar	516
Romania	303

Nation	Number of students
Russian Federation	320
Saudi Arabia	307
Scotland	290
Serbia	288
Singapore	328
Slovenia	292
Sweden	369
Syria, Arab Republic of	327
Thailand	390
Tunisia	292
Turkey	314
Ukraine	321
United States	544

***Appendix B: Listing of the National Level Curriculum Explanatory Variables***

Variable	Description	Data coding
1. Calculator	Does the national curriculum contain statements/policies about the use of calculators in grade 8 mathematics?	Binary 0/1; Yes = 1
2. Computer	Does the national curriculum contain statements/policies about the use of computers in grade 8 mathematics?	Binary 0/1; Yes = 1
<i>How much emphasis does the national mathematics curriculum place on the following?</i>		
3a. Basic	(a) Mastering basic skills and procedures	4 point scale; None = 0, Very Little = 1, Some = 2, A lot = 3
3b. Concept	(b) Understanding mathematical concepts and principles	4 point scale; None = 0, Very Little = 1, Some = 2, A lot = 3
3c. Real life	(c) Applying mathematics in real-life contexts	4 point scale; None = 0, Very Little = 1, Some = 2, A lot = 3

(continued)

Variable	Description	Data coding
3d. Communicate	(d) Communicating mathematically	4 point scale; None = 0, Very Little = 1, Some = 2, A lot = 3
3e. Reason	(e) Reasoning mathematically	4 point scale; None = 0, Very Little = 1, Some = 2, A lot = 3
3f. Integrating	(f) Integrating mathematics with other subjects	4 point scale; None = 0, Very Little = 1, Some = 2, A lot = 3
3g. Proof	(g) Deriving formal proofs	4 point scale; None = 0, Very Little = 1, Some = 2, A lot = 3
4a & b. Which best describes how the mathematics curriculum addresses the issue of students with different levels of ability? (Two variables DFlevel and DFcur)	Different curricula are prescribed for students of different ability levels.	Design Matrix DFlevel      DFcur 0   1
	The same curriculum is prescribed for students of different ability levels, but at different levels of difficulty	1   0
	The same curriculum is prescribed for all students	0   0
5. Remedial	Is there an official policy to provide remedial mathematics instruction at the eighth grade of formal schooling?	Binary 0/1; Yes = 1
6. Degree	Which are the current requirements for being a middle/lower secondary grade teacher? A degree from a teacher education program	Binary 0/1; Yes = 1
7. Exam	Across grades K-12, does an education authority in your country (e.g., National Ministry of Education) administer examinations in mathematics that have consequences for individual students, such as determining grade promotion, entry to a higher school system, entry to a university, and/or exiting or graduating from high school?	Binary 0/1; Yes = 1



## References

- Chen, G., Mathieu, J. E., & Bliese, P. D. (2004a). A framework for conducting multilevel construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Research in multilevel issues: Multilevel issues in organizational behavior and processes* (Vol. 3, pp. 273–303). Oxford, UK: Elsevier.
- Chen, G., Mathieu, J. E., & Bliese, P. D. (2004b). Validating frogs and ponds in multilevel contexts: Some afterthoughts. In F. J. Yammarino & F. Dansereau (Eds.), *Research in multilevel issues: Multilevel issues in organizational behavior and processes* (Vol. 3, pp. 335–343). Oxford, UK: Elsevier.
- Dansereau, F., & Yammarino, F. J. (2000). Within and between analysis: The variant paradigm as an underlying approach to theory building and testing. In K. J. Klein & S. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 425–466). San Francisco, CA: Jossey-Bass.
- Forer, B., & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research: An International Interdisciplinary Journal for Quality of Life Measurement*, *103*, 231–265. doi:10.1007/s11205-011-9844-3.
- Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, *53*, 455–467.
- Hofmann, D. A., & Jones, L.M. (2004). Some foundational and guiding questions for multilevel construct validation. In F. Yammarino & F. Dansereau (Eds.), *Multi-level issues in organizational behavior and processes*. Amsterdam: Elsevier.
- Kaplan, D., & Elliott, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling*, *4*, 1–24.
- Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review*, *19*, 195–229.
- Lee, S.-Y. (1990). Multilevel analysis of structural equation models. *Biometrika*, *77*, 763–772.
- Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika*, *57*, 581–597.
- Morgeson, F. P., & Hofmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review*, *24*, 249–265.
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. URL: [http://timss.bc.edu/timss2007/PDF/T07\\_AF.pdf](http://timss.bc.edu/timss2007/PDF/T07_AF.pdf).
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398.
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316.
- Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, *16*, 295–330.
- Stone, J., & Zumbo, B. D. (2016). Validity as a Pragmatist project: A global concern with local application. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice* (pp. 555–573). Newcastle, UK: Cambridge Scholars Publishing.
- Watkins, K. (2007). *Human development report 2007/2008, fighting climate change: Human solidarity in a divided world*. New York, NY: United Nations Development Programme.
- Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45–79). Amsterdam, The Netherlands: 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, NC: IAP – Information Age Publishing, Inc..

- Zumbo, B. D., & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High stakes testing in education – Science and practice in K-12 settings* (pp. 177–190). Washington, DC: American Psychological Association Press.
- Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, 5, 1–23.
- Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Astivia, O. L. O., & Ark, T. K. (2015). A methodology for Zumbo's Third Generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, 12, 136–151.

# Chapter 19

## On Models and Modeling in Measurement and Validation Studies

**Bruno D. Zumbo**

This chapter is motivated by three observations. First, as is evident from the collection of chapters in this edited volume, models and the process of modeling are growing in importance and centrality in the theoretical and empirical analyses of the response processes evidence for measurement validity (e.g., Launeanu & Hubley, Chaps. 6 and 7, this volume). Models and modeling are key ingredients that bring to life the *Standards'* (AERA, APA, NCME, 2014) statement: “Inferences about processes involved in performance can also be developed by analyzing the relationship among parts of the test and between the test and other variables” (p. 15). For example, Zumbo et al. (Chap. 18, this volume) and Chen and Zumbo (Chap. 4, this volume) illustrate the use of an ecological model to move beyond the traditional focus on individual test-takers' behaviors to an explanation-focused view of validity, and hence item responding. The combination of an ecological model of item and test responding and the explanation-focused view of validity bridges the inferential gap from the test data to response processes and provides inferential strength to the conclusions based on the empirical data modeling. It should be noted that by ‘inferential strength’ I mean the amount of support that the evidence or reasons provide the conclusion about response processes (and hence validity); and is therefore considered a matter of degree such that the more support (the more evidence or reasons) there is for a conclusion, the stronger the argument for the conclusion.

The second motivating factor for this chapter is that in contemporary measurement and validation practices, which are heavily model-based, the inferences, in part, arise from and are supported by the model itself. In short, the statements about the validity of the inferences from the test scores rest on the measurement model.

---

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: [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca)

In fact, given that in the process of empirical modeling one, in essence, begins with an array of numbers denoting responses to items or tasks for each examinee, it could be argued that the psychometric model “provides” the inferences one can make by being the vehicle for going from what we have to what we wish we had – that is, we have item or task responses but, as examples, we wish we had the process of item responding or, the score on the latent variable being measured by the test. The upshot is that the measurement model is not neutral in the validation (and more generally the measurement) process because it helps us travel from the item responses to the test takers’ response processes and/or their status on the latent variable of interest. Therefore, not surprisingly, one’s test score interpretations may change depending on the psychometric statistical model being used. As Zumbo (2007) and Hubley and Zumbo (2013) note, discussions of ‘validity’ and ‘validation’ are nearly always framed and shaped by the (measurement and psychometric) models employed, be they classical or observed-score test theory, item response theory, factor analysis, or axiomatic scaling theory. Therefore, measurement models are not neutral in the validation process because they have (or bring to the modeling process) their own underlying values and assumptions and their consideration is necessary for a fulsome discussion of validity (Zumbo, 2007).

This chapter follows closely some of my earlier work (Zumbo, 2007; Zumbo & MacMillan, 1999) and therefore the third motivation for this chapter is that given that the central message about the role of models has mostly gone unnoticed, or unheard, it bears repeating here. In the remainder of this chapter I first make some remarks about how the term ‘model’ is used and its implications for measurement practice and validation.

## Remarks About How the Term Model Is Used

In measurement and validation research, like all research, often we use the term ‘model’ to convey the sense of a mathematical model; a model in the wider philosophical sense; a model in the psychological sense of a psychological model of a phenomenon (e.g., a psychological model of stress); an explanatory model; a descriptive model; a stochastic or random variable model; a statistical model, a logical model; and a computational model, to list but a few. In day-to-day measurement and validation practices what complicates matters is that these uses of ‘model’ are not mutually exclusive (nor exhaustive) but they do have essential but subtle distinctions. In what follows, as I describe the various defining features of models, I will shed some light on psychometric models and their use in response processes and more general validation practices.

In the technical philosophy of science literature there are two distinct meanings of ‘model’: postulational or axiomatic and iconic. In certain formal disciplines such as logic and mathematics a model (for or of a theory) has its roots in the axiomatic or postulational method of deductive systems including some branches of modern mathematics. The basic idea of the axiomatic method is that the content of a

scientific subject should consist of a set of assumed propositions, called axioms or postulates, and that other propositions, called theorems, should be derived from the basic assumptions by applying the rules of deductive logic. Note that the axioms must be accepted without proof. However, if the scientific subject under consideration in this axiomatic or postulational-deductive fashion is to be practical, realistic and purposeful, the axioms are usually selected so as to approximate or idealize actual experience. The attitude of pure or abstract modern mathematics is quite different from the one just described. In it one has the right to choose the content of axioms somewhat arbitrarily (often allowing for undefined terms or empty symbols), subject only to certain logical criteria such as consistency. A set of postulates is said to be consistent if there exists an interpretation of the undefined terms which converts all the postulates into true statements. In mathematical logic the result of such interpretation, that is, the concrete set of true statements, is called a 'model' of the abstract postulate system. In so doing the abstract deductive system is said to be transformed from an abstract theory into a concrete theory.

Therefore, two models, in this mathematical sense, can then be quite different. For example, Kolmogorov laid the set-theoretic foundation for a probability theory for which there are two interpretations (or models) of probability: Bayesian and Frequentist. Similarly, quantum and relativistic physics have a variety of interpretations. Likewise one can ask whether the Fisher and Neyman-Pearson theories of testing hypotheses are one model or two separate models (Lehmann, 1993). In the context of psychometric models, this is not unlike the two interpretations of item response theory as a Rasch model versus item response modeling via 1 through 4 parameter logistic item response models. Finally, as a last example in measurement, one has the "stochastic subject" versus the random sampling interpretations of the probability in item response theory models (Holland, 1990a).

In the history of psychometric measurement this sort of modeling and casting of axioms and interpretations is most clearly seen in the varied formalizations of the principal results of test theory by Thurstone (1932), Gulliksen (1950), Guttman (see Zimmerman, Williams, Zumbo, & Ross, 2005), Novick (1966), Lord and Novick (1968), Rozeboom (1966), Holland (1990b), and Zimmerman and Zumbo (2001) wherein one sees the various axioms and principal results of psychometric and test theories from statistical, algebraic, probabilistic, and geometric mathematical perspectives. Applying an observation by Zumbo and Kroc (2016), there is a distinction between probability in practice (the ultimate subject of psychometrics) and the platonic structure of the mathematical objects that we use to conveniently describe that practice. These descriptions are nearly always approximations: we simplify our practical probability spaces by smudging them into theoretical ones. This has undeniably proven to be an extremely fruitful tactic, but it has also given rise to several conundrums and apparent paradoxes. Progress is often made at this level of abstraction by challenging the needs for certain axioms, restating axioms, or resolving apparent paradoxes arising from the mathematical logic. The primary concern of psychometric models and modeling techniques is to give us a clearer idea of the psychological phenomenon (also sometimes called a trait or latent variable) we wish to measure, its relationship to similar phenomenon, and the test circumstances

that are most likely to provide a valid measure of the quantities in which it exists in various situations.

The second meaning of ‘model’ in the technical philosophy of science literature involves analogues of things and/or processes. Some real or imagined thing, or process, behaves similarly to some other thing or process. In this sense a mathematical model is an abstract idealization of various features of a real situation in the same sense that pure Euclidean plane geometry is the abstract counterpart of the surveyors’ concept of physical points, lines, polygons, circles, etc. and their properties.

It should be noted that although we consistently refer to mathematical models, the term ‘model’ has a wider-sense meaning in the philosophies of science. For example, we have the molecule model of gas, the constitution and cataloging of personality traits and variables similar to a periodic table of elements in chemistry, or the computer model of human cognition or memory. Furthermore, models are used for certain definite purposes in the sciences. For example, they enable certain inferences to be made which would not otherwise be possible. This is a logical purpose. In addition, they may serve an epistemological purpose by expressing our knowledge of the world, and enabling us to delineate and extend our knowledge of the world.

Finally, it is instructive to note that models can be homeomorphs or paramorphs. The essential difference between these is the source of a model and the subject of a model. For example, a model airplane has as its source the real thing, the airplane, while its subject is the airplane—a homeomorph. On the other hand, when one is using the computer as a model for human information processing and cognition, the computer is not modeled on cognition in any way at all. The computer is modeled on something quite different, namely principles in logic and solid state physics; hence a paramorph for cognition and information processing. The two main uses of models in science are heuristic, to simplify a phenomenon; or explanatory, to for example describe the causal mechanism which produces the phenomena. A case has been made in the philosophies of science literature that explanatory models are, or use, paramorphic models.

### ***What Does This all Mean for Measurement Practices and Validation?***

First, all measurement models known to me are paramorphic but heuristic. This lack of explanatory focus has been the root of a long-standing anxiety among some measurement specialists (e.g., Zumbo, 2009). Historically, attempts at relieving this anxiety has been to prevail on cognitive theory to lend an explanatory hand; noting, however, that not all cognitive theories are explanatory so that we need to be careful that we do not inadvertently supplant one heuristic model with another while deluding ourselves that our new model is explanatory.

Second, most measurement models vary on the degree of iconic and/or axiomatic or postulational focus. For example, there exists an extensive literature on purely axiomatic measurement. Such models of measurement have been used almost exclusively in psychophysics, decision sciences and mathematical social/behavioral sciences. These models, however, are not used in the everyday practice of validation or measurement, nor were they necessarily intended to be. The major cause of this lack of use of the models inspired by psychophysics is that they are mostly deterministic models so that one is left asking how many axioms need to be false before a model is not useable. In this light, the most common result discussed outside of the fields of psychophysics and decision sciences is that of scaling theory and scales of measurement. In this light, the Rasch model has some kinship with this axiomatic approach. One could make a case that the Rasch model is a probabilistic/stochastic variant one of these traditional deterministic models, additive conjoint measurement.

Most of the psychometric models in practice today have both a deterministic (or structural) component and a stochastic component, and most are of the iconic variety for example item response theory and generalizability theory. The purpose of these models is to allow us to make certain inferences about test scores. In this light, the Rasch model has an essential difference (and one that distinguishes it from the one-parameter logistic model) in that it also has an epistemological purpose. However, for simple inferential purposes the Rasch model has much in common with item response theory, but its epistemological purpose sets it aside. This epistemological purpose is, beyond a doubt, controversial because some psychometricians may argue that psychometric models should not serve epistemological purposes but rather should only aid in inference. As Goldstein and Wood (1989) highlight (see Hubley et al., Chap. 5, this volume), current widespread uses of item response theory are focused on statistical estimation and use of the item and person parameters in test assembly, equating/linking, and test scoring. In this case, the model is a practical tool to aid day-to-day tasks. As Goldstein and Wood state, this is certainly fine for operational testing purposes, hence making it mostly an exercise in statistical modeling, but this practice ignores item response theory's history in twentieth century psychological theorizing – de-emphasizing its explanatory value and its connection to response processes.

Finally, axiomatic models do not rely on evidence from measurement validity theory for buttressing their score inferences but rather on proofs of uniqueness and representational theorems. The iconic models, on the other hand, like classical test theory, item response theory and generalizability theory rely on validity theory for validating their inferences. Again, in this light the Rasch model relies on its axiomatic and epistemological kinship to aid in its interpretational framework (i.e., the use of interval scale measurement).



## **Concluding Remarks: Measurement Modeling Practices and Model Choice**

Models and modeling have a profound impact on the ways in which knowledge, aptitudes, competencies and psycho-social characteristics are assessed and even conceptualized. Along with rapid technological change to new and more varied psychometric and statistical models in measurement practice writ large, and validation practices in particular, will come an increased premium on appropriate model usage and interpretation. Because a one-model fits all approach is unsuited to measurement practices and validation one kind of model is unlikely to be seriously endorsed for long.

There are several reasons for the great appeal of models and modeling to researchers, policymakers, and measurement specialists. First, with the advent of cheap and easily attained computers and easily used statistical software, modeling is relatively inexpensive. Second, the choice of models can be externally mandated. It is far easier to mandate what one uses (either directly through policy by a governing body or indirectly through trends and fads in model choice as well as local organizational and scholarly/community customs) than it is to choose models based on a comprehensive analysis of empirical model adequacy and fit. Third, change in model usage can be rapidly implemented so no model commitment is a life-time commitment. Fourth, modeling results are visible. That is, the results of modeling such as tests of fit and reporting of estimated parameter results can be easily reported to an interested reader. Of course, models and modeling practices come in many different forms and may be used in a variety of ways in measurement research and validation practices.

The last reason for the great appeal of measurement models is that in practice no matter how much data you have; it is never enough because without complete information you will always have some error of measurement or fallible indicator variable. The function of the psychometric model in measurement and validity research is to step in when the data are incomplete. In an important sense, we are going from what we have to what we wish we had. If we had available the complete data or information, then we would know the true score, or theta in IRT models, and no statistics beyond simple summaries would be required. There would be no need for complex models to infer the unobserved score from the observed data and, hence, no need to check the adequacy and appropriateness of such inferences through validation. We get around data and information limitations by augmenting our data with assumptions. In practice, we are, in essence, using the statistical model to create new data to replace the inadequate data. For example, the most common data augmentation assumption in psychometrics is that the dependencies (e.g., correlations) among items are accounted for by an unobserved continuum of variation – of prominence in item response theory and factor analysis models.

At this point a word of caution is important. In Greek mythology Pygmalion fell in love with one of his sculptures, which then came to life. Many of us are like the fictional mythical character Pygmalion and fall in love with our models; in good

part we fall in love with what we want our model be. In some cases we are very much like Pygmalion in that we believe that our particular model of interest becomes real – through our love we make it real. The danger in this type of magical thinking, however, is that in our cases these (psychometric and measurement) models are used in influential and high-stakes decision making. This type of behavior inspired by love for our particular model, and magical thinking, results in psychometric model use as a kind of ritualistic cultural behavior that continues unabated because these model choices and practices appear (at least to some measurement practitioners) to be objective and exact, they are easily and readily available in statistical software packages, students are taught to use them, and journal reviewers and editors demand them. To be clear, in these remarks I am not thinking of any one particular measurement practice (e.g., Rasch modeling) even if that practice corresponds to my description, but rather all model choice and practices – for example, I am of the vintage to remember all too well the time of the *LISREL-ites* in the desert of social and behavioral research casually making causal claims.

Therefore, in conclusion, one of my central messages in this chapter is that not only are there a variety of models and modeling practices in scientific practice, but that models are empirical commitments. This point is best made by Zumbo and Rupp (2004):

*It is the responsibility of mathematically trained psychometricians to inform those who are less versed in the theory about the consequences of their decisions to ensure that examinees are assessed fairly. Because models (which, in part, include the parameter estimation strategy) are empirical commitments, it is measurement specialists who need to take partial responsibility for the decisions that are being made with the models they provide to others. Everyone knows that a useful and essential tool such as an automobile, a chainsaw, or a statistical model can be very dangerous if put into the hands of people who do not have sufficient training and handling experience or lack the willingness to be responsible users. (p. 87)*

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology*, *42*, 139–167.
- Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.
- Holland, P. W. (1990a). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601.
- Holland, P. W. (1990b). The Dutch Identity: A new tool for the study of item response models. *Psychometrika*, *55*, 5–18.
- 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.

- Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88, 1242–1249.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3, 1–18.
- Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood, IL: The Dorsey Press.
- Thurstone, L. L. (1932). *The reliability and validity of tests*. Ann Arbor, MI: Author.
- Zimmerman, D. W., Williams, R. H., Zumbo, B. D., & Ross, D. (2005). Louis Guttman's contributions to classical test theory. *International Journal of Testing*, 5, 81–95.
- Zimmerman, D. W., & Zumbo, B. D. (2001). The geometry of probability, statistics, and test theory. *International Journal of Testing*, 1, 283–303.
- Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45–79). Amsterdam, The Netherlands: 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, NC: IAP – Information Age Publishing, Inc..
- Zumbo, B. D., & Kroc, E. (2016). Some remarks on Rao and Lovric's testing the point null hypothesis of a normal mean and the truth: 21st century perspective. *Journal of Modern Applied Statistical Methods*, 15, 33–40.
- Zumbo, B. D., & MacMillan, P. O. (1999). An overview and some observations on the psychometric models used in computer-adaptive language testing. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 216–228). Cambridge, UK: Cambridge University Press.
- Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73–92). Thousand Oaks, CA: Sage Press.