Quan Zhang *Editor*

Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings

Springer

Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings

Quan Zhang
Editor

Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings

Springer

Editor
Quan Zhang
College of Foreign Studies
Jiaxing University
Jiaxing, Zhejiang
China

ISBN 978-981-10-1686-8 ISBN 978-981-10-1687-5 (eBook)
DOI 10.1007/978-981-10-1687-5

Library of Congress Control Number: 2016943454

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media Singapore Pte Ltd.

Preface

PROMS 2015 symposium was held on campus of Kyushu Sangyo University, Fukuoka, Japan from August 20 to 24, 2015 with pre-conference workshops scheduled for August 20–21, 2015. And the present conference proceedings were done ad hoc for the researches, Ph.D. supervisors, educators, practitioners and younger generation who seek to use the Rasch model in their research activities in Pacific Rim countries, regions and beyond.

More than half a century has passed since the Danish mathematician Georg Rasch (1901–1981) published his "Probabilistic Model for Intelligence and Attainment Tests" (Rasch 1960). With this departure, the model has found wide applications in measuring variables ranging from business, counseling, economics, education, health care, language testing, measurement, psychology, quality assurance, statistics to strategic planning field and has been extended from the initial application to the dichotomous data type to the polytomous ones. Today, the model is held as "Rasch Model" among measurement professionals and believed to have instigated the vision of promoting objective measurement and to have contributed greatly to scientific discovery.

To this end, Pacific Rim Objective Measurement Symposium (PROMS) has been devoting all their endeavors over the past decade and PROMS conferences have been successfully hosted in many Pacific Rim countries and regions for such a purpose of promoting the research and contributing to the development of the Rasch Model. PROMS 2015 Fukuoka, Japan is the eleventh symposium and follows highly successful meetings in PROMS 2005 Kuala Lumpur, PROMS 2006 Hong Kong, PROMS 2007 Taiwan, PROMS 2008 Tokyo, PROMS 2009 Hong Kong, PROMS 2010 Kuala Lumpur, PROMS 2011 Singapore, PROMS Jiaxing, China Mainland, and PROMS 2013 Kaohsiung, Taiwan, PROMS 2014, Guangzhou, China Mainland, and PROMS 2016 is to be held in Xi'an, China Mainland again. Just as Prof. Rob Cavanagh, the Chair of PROMS, said, there are many good reasons why you should attend PROMS: "The keynote speakers and workshop presenters are all eminent scientists with cutting-edge expertise in Rasch measurement and its applications; students are encouraged to attend and present

their work. The atmosphere is highly collegial and we value the contributions of all; PROMS is highly supportive of early career researchers and the professors who attend are renowned for the support they provide for all participants". Therefore, anyone seriously interested in research and development of psychometrics or measurement will find such international symposiums and related workshops to be an excellent source of information about the application of the Rasch Model.

The present volume contains 27 articles germane to Rasch-based research work submitted by scholars from PROMS 2015. Each of these articles deals with Rasch measures in their research field, covering a variety of issues ranging from education, psychology, management, language testing to medicine and serving in particular as good resources for researchers and students to be able to conduct their own Rasch model analyses as well as understand and review published Rasch-based research.

Our sincere thanks go to all the contributors for their time and efforts to make this book a reality. And we should thank the Springer for the publication of this book; we also express our sincere gratitude to the MetaMetrics, USA, who has been acting as a great sponsor for PROMS each year over the past years and statistics.com as well.

Apart from this, we also want to thank and appreciate the proofreading carefully done by Rasch peers. While they save us from a number of inaccuracies and infelicities, they can in no way be held responsible for the academic opinions which are expressed and the imperfection which no doubt remains. And in particular, great thanks should go to Dr. Durand, Jeff, the newly appointed deputy chair from Toyo Gakuen University, Japan for all the efforts and time he devoted before and during the editing of the present book.

Copies of the present book will be sent to universities and colleges in Pacific Rim countries and regions as well as Europe and other parts of the world. In doing so, we are confident to claim that the past decade of PROMS has been rewarding. The retrospect is impressive and the prospect is tantalizing. With updated computer technology and development, the time for objective measurement via Rasch has come of age.

Jiaxing University, China Quan Zhang

Reference

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. Wright, University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

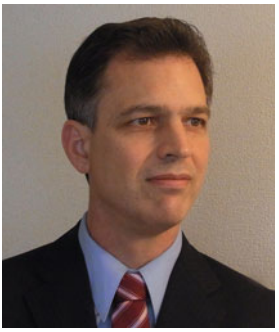
PROMS Board Members (2015/2016)

The PROMS Board of Management comprises academics and researchers who oversee the maintenance of the culture and traditions of PROMS. This includes championing the application and dissemination of the Rasch model and modern measurement theory. In addition to the chair and deputy chair, members include advisors and contributors invited by the board, one representative from each country or region involved in PROMS, and members of the organizing committee for the next symposium.

Chair
Prof. Cavanagh, Rob, Curtin University, Australia

Past Chair
Prof. Bond, Trevor, James Cook University, Australia

Deputy Chair
Dr. Yan, Zi, The Educational University of HK, Hong Kong SAR



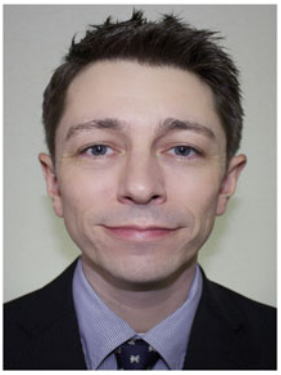
2nd Deputy Chair
Dr. Durand, Jeff, Toyo Gakuen University, Japan

Secretary/Mainland Chinese Delegate
Prof. Zhang, Quan, Jiaxing University, China

Malaysian Delegate
Dr. Mohd Nor, Mohd Zali, Newstar Agencies Sdn Bhd, Malaysia

USA Delegate
Prof. Engelhard, George, University of Georgia, USA

Singaporean Delegate
Dr. Lee, Iris, Ministry of Education, Singapore



Web Administrator/Japanese Delegate
Dr. Batty, Aaron, Keio University, Japan

Hong Kong Delegate
Prof. Wang Wen Chung, The Educational University of HK, Hong Kong, SAR

Board Advisors

Prof. Stenner, Jackson, MetaMetrics, Inc., USA

Dr. Fisher, William P. Jr., Living Capital Metrics, USA

PROMS 2015 Pre-conference Workshops

In accordance with the international conference practice, each PROMS program is preceded by two days of workshops. These typically provide research training on: the basics of Rasch measurement using Winsteps; measuring English language performance with Lexiles; many-facets Rasch measurement using Facets; computer-adaptive testing; evaluating the quality of performance assessments; constructing measures; and multi-dimensional Rasch models. Listed below are six pre-conference workshops conducted from August 20 to 24, 2015 in Fukuoka, Japan.

Pre-conference Workshop I. August 20, 2015, 9:00 am–4:30 pm

Item banking using the Rasch Measurement Model (1-day) by Dr. Lead Psychometrician, Rassoil Sadeghi, Australian Curriculum, Assessment and Reporting Authority (ACARA). The workshop focuses on the following topics:

1. Rasch Measurement model with an emphasis on its distinctive features;
2. Item banking: applications, advantages and limitations;
3. Item banking using the Rasch measurement model 1, model 2; and
4. Application of item bank in school assessments

RUMM software was used to show how an item bank can be created using the Rasch measurement model.

Pre-conference Workshop II. August 20–21, 2015, 9:00 am–4:30 pm

An introduction to Rasch analysis using Winsteps (2-days) conducted by Prof. Trevor Bond. provides hands-on experience. Professor Bond introduced the rationale for using the Rasch model, and the participants worked through a series of guided hands-on data analysis exercises. The workshop focused on the dichotomous Rasch model and the function of fit indices. The application of the Rating scale model to Likert-style data was introduced. Latest Rasch software was available for all participants to download. Tutorial worksheets were available in English/Japanese.

Pre-conference Workshop III. August 20, 2015, :00 am–12:30 pm

This workshop was conducted by Prof. Jackson Stenner who introduced the concept of Causal Rasch models (half a day). All measurements share a three-part structure: (1) an attribute such as human temperature or reading ability; (2) a measurement mechanism that transmits variation in the attribute to; (3) a measurement outcome such as a count correct on a reading test or a count of cavities turning black on a Nextemp™ thermometer. Causal Rasch models expose the measurement mechanism and enable direct tests of competing construct theories.

This workshop introduces applications over a wide range of attributes including reading ability, mathematical ability, and short-term memory. Participants were encouraged to read Causal Rasch Models by Stenner et al. (2013) in Frontiers in Psychology. The workshop format was informal and discussions were encouraged.

Pre-conference Workshop IV. August 20, 2015, 1:30 pm–4:30 pm

This half-day workshop was conducted by James Sick who provided a detailed introduction to the many-facet Rasch model by exploring features of the Facets software package. Although a general knowledge of Rasch measurement is assumed, a brief overview of the many-facet model and its applications was presented. The session covered:

Formatting data for a Facets analysis,
Writing a basic Facets control file,
Interpreting standard Facets output, including vertical rulers, measurement tables, and category probability charts.

A time-limited edition of Facets was provided to all workshop participants. Facets is a Windows only program. To use Facets, participants need to bring a Windows laptop, or a Mac equipped to run Windows. Windows programs can be run on Macintosh computers after installing Windows via Apple Bootcamp, or with emulation software such as Parallels Desktop or VMware Fusion.

Pre-conference Workshop V. August 21, 2015, 9:00 pm–4:30 pm

This workshop was conducted by Prof. Tetsuo Kimura who provided participants with an overview of basic concepts of computer-adaptive testing (CAT) and an opportunity to implement a Rasch-based small-scale CAT with open-source software. After briefly discussing key concepts in CAT such as item banking, item selection, target difficulty, maximum information, stopping rules and standard error, two Rasch-based CAT programs were introduced: UCAT (Linacre 1987) and Moodle UCAT (Kimura et al. 2012). Participants were encouraged to bring their own notebook PCs so that they can practice building an item bank and creating CATs in the open-source learning management system Moodle. A temporary teacher account on the Moodle UCAT server was given to each participant.
The workshop was divided into four sessions as follows:

Session 1: Computer-adaptive testing (CAT)—its origins and concepts
Session 2: UCAT and Moodle UCAT

Session 3: Building an item bank on open-source LMS Moodle
Session 4: Creating CATs on Moodle UCAT

Pre-conference Workshop VI. August 21, 2015, 9:00 pm–4:30 pm

This workshop was conducted jointly by Prof. George Engelhard Jr. and Jue Wang to deal with invariant measurement with raters and rating scales. According to the workshop runners, the use of rating scales by raters is a popular approach for collecting human judgments in numerous situations. In fact, it is fair to say that raters and rating scales in the social, behavioral and health sciences are ubiquitous. Raters appear in applied settings that range from high-stakes performance assessments in education through personnel evaluations in a variety of occupations to functional assessments in medical research. This workshop utilized the principles of invariant measurement (Engelhard 2013) combined with lens models from cognitive psychology to examine judgmental processes that arise in rater-mediated assessments. This workshop focused on guiding principles that can be used for the creation, evaluation and maintenance of invariant assessment systems based on human judgments.

The purpose of this workshop was to provide an introduction to the concept of invariant measurement for rater-mediated assessments, such as performance assessments. Rasch models provide an approach for creating item-invariant person measurement and person-invariant item calibration. This workshop extended these ideas to measurement situations that require raters to make judgments regarding performance assessments. This workshop provided an introduction to the many-facet model, and its use in the development of psychometrically sound performance assessments. Examples were based on large-scale writing assessments. Participants were encouraged to bring their own data sets for analysis and discussion in the workshop.

The Facets computer program (Linacre 2007) was used throughout the workshop to illustrate the principles of invariant measurement with raters and rating scales. Reference www.GeorgeEngelhard.com.

PROMS 2015 Local Committee

The Local Committee, headed by Aaron Batty and Jeff Stewart, comprised a team of student volunteers who did the actual work, i.e., conference promotion, implementation and organizational details, conference budget and so forth.

Aaron Batty and Jeff Stewart (Chairs)

Members:

Jeffrey Durand,
James Sick,
Paul Horness,
Stuart McLean

And student volunteers of Kyushu Sangyo University, Fukuoka, Japan

Sponsors and Donors for PROMS 2015, Fukuoka, Japan

Editorial Board Members

Anonymous peer reviewers from Universities and Institutes of Pacific-Rim and beyond

Acknowledgements

The following individuals and sponsors helped to make the Pacific Rim Objective Measurement Symposium (PROMS), held in Fukuoka, Japan (August 20–24, 2015), a success. Their contributions in maintaining the quality of paper presentation, review, running workshops and the organization of the international academic conference are greatly appreciated.

Prof. Cavanagh, Robert (Chair), Curtin University, Australia
Prof. Bond, Trevor (Past Chair), James Cook University, Australia
Dr. Durand, Jeff (Vice Chair), Toyo Gakuen University, Japan
Prof. Zhang, Quan (Secretary), University of Jiaxing, China
Prof. Stenner, Jackson, MetaMetrics, Inc., USA
Prof. Engelhard, George, University of Georgia, USA
Dr. Lee, Iris, National Institute of Education, Singapore
Prof. Wang, Wen Chung, The Educational University of Hong Kong, Hong Kong, SAR
Dr. Mohd Zali Mohd Nor, Newstar Agencies Sdn Bhd, Malaysia
Dr. Batty, Aaron, Keio University, Japan
Dr. Jue Wang, University of Georgia Educational Psychology, USA
Dr. Rassoil Sadeghi, Australian Curriculum, Assessment and Reporting Authority (ACARA), Australia
Prof. James Sick, International Christian University, Tokyo, Japan
Prof. Tetsuo Kimura, Niigata Seiryo University, Japan

And student volunteers of Kyushu Sangyo University, Fukuoka, Japan.

Contents

Causal Rasch Models in Language Testing: An Application Rich Primer

A. Jackson Stenner, Mark Stone, William P. Fisher Jr. and Donald Burdick

A new paradigm for measurement in education and psychology, which mimics much more closely what goes on in the physical sciences was foreshadowed by Thurstone (1926) and Rasch (1961):

> It should be possible to omit several test questions at different levels of the scale without affecting the individual's score [measure].
>
> … a comparison between two individuals should be independent of which stimuli [test questions] within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion.

Taken to the extreme, we can imagine a group of language test takers (reading, writing, speaking, or listening) being invariantly located on a scale without sharing a single item in common. i.e. no item is taken by more than one person. This context defines the limit case of omitting items and making comparisons independent of the particular questions answered by any test taker.

More formally we can contrast a fully crossed pxi design (persons crossed with items) in which all persons take the same set of items with a nested design i:p (all items are unique to a specific person). The more common design in language research is pxi simply because there is no method of data analysis that can extract

A.J. Stenner (✉) · D. Burdick
MetaMetrics, Durham, NC, USA
e-mail: jstenner@lexile.com

A.J. Stenner
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

M. Stone
Aurora University, Aurora, IL, USA

W.P. Fisher Jr.
University of California—Berkeley, Berkeley, CA, USA

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_1

invariant comparisons from an i:p design unless item calibrations are available from a previous calibration study or are theoretically specified.

But the i:p design is routinely encountered in physical science measurement contexts and in health care when, for example, parents report their child's temperature to a pediatrician. Children in different families do not share the same thermometers. Furthermore, the thermometers may not even share the same measurement mechanism (mercury in a tube vs. NexTemp technology, see Note 1). Yet, there is little doubt that the children can be invariantly ordered and spaced on any of several temperature scales.

The difference between the typical language testing and temperature scenarios is that the same construct theory, engineering specifications and manufacturing quality control procedures have been enforced for each and every thermometer, even though the measurement mechanism may vary. In addition, considerable resources have been expended in ensuring the measuring unit (°F or °C) has been consistently mapped to the measurement outcome (e.g. column height of mercury or cavity count turning black on a NexTemp thermometer) (Hunter 1980; Latour 1987). Substantive theory, engineering specifications, and functioning metrological networks—not data- render comparable measurement from these disparate thermometers. This contrast illustrates the dominant distinguishing feature between measurement in the physical and educational sciences including EFL, ESL and ENL language testing. Educational measurement does not, as a rule, make use of substantive theory in the ways the physical sciences do (Taagepera 2008). Nor does educational science embrace metric unification even when constructs (e.g. reading ability) repeatedly assert their separate independent existences (Fisher 1997, 1999, 2000a, b; Fisher et al. 1995).

Typical applications of Rasch models in language testing are thin on substantive theory. Rarely is there an a priori specification of the item calibrations (i.e. constrained model). Instead the researcher estimates both person parameters and item parameters from the same pxi data set. For Kuhn (1961) this practice is at odds with the scientific function of measurement in that substantive theory almost never will be revealed by measuring. Rather "the scientist often seems to be struggling with facts [measurement outcomes, raw scores], trying to force them to conformity with a theory s(he) does not doubt" (p. 163). Kuhn is speaking about substantive construct theory, not axiomatic measurement theory. Demonstrating data fit to a descriptive Rasch Model or sculpting a data set by eliminating misfitting items and persons and then rerunning the Rasch analysis to achieve satisfactory fit is, specifically not, the "struggling" Kuhn is referring to.

The gold standard demonstration that a construct is well specified is the capability to manufacture strictly parallel instruments. A strictly parallel instrument is one in which the correspondence table linking attribute measure to measurement outcome (count correct) is identical although items are different on each parallel instrument. So, imagine two 4000 word 1300L articles, one on 'atomic theory' and one on 'mythology'. Both articles are submitted to a machine that builds 45 four choice cloze items distributed about one item for every 80–100 words. These one-off items are assumed to have calibrations sampled from a normal distribution with a mean

equal to 1300L and a standard deviation equal to 132L. With this information, an ensemble Rasch model (Lattanzio et al. 2012) can produce a correspondence table linking count correct to Lexile measure. Since the specifications (test length, text measure, text length and item spread) are identical for the two articles, the correspondence tables will also be identical; on both forms 25 correct answers converts to 1151L and 40 correct answers converts to 1513L, and so on. The same basic structure plays out with NexTemp® thermometers. A NexTemp® thermometer has 45 cavities. Twenty-five cavities turning black converts to a temperature of 37.9 °C, whereas 40 cavities turning black converts to 39.4 °C. In both cases theory, engineering specifications and manufacturing guidelines combine to produce strictly parallel instruments for measuring reading ability and human temperature and in each case it is possible to manufacture large quantities of identical instruments. The capacity to manufacture "strictly" parallel instruments is a milestone in an evolving understanding of an attribute and its measurement. Richard Feynman wrote: "What I cannot create, I don't understand!" We demonstrate our understanding of how an instrument works by creating copies that function like the original.

Descriptive Rasch Models Versus Causal Rasch Models

Andrich (2004) makes the case that Rasch models are powerful tools precisely because they are prescriptive, not descriptive, and when model prescriptions meet data, anomalies arise. Rasch models invert the traditional statistical data-model relationship. Rasch models state a set of requirements that data must meet if those data are to be useful in making measurements. These model requirements are independent of the data. It does not matter if the data are bar presses, counts correct on a reading test, or wine taste preferences, if these data are to be useful in making measures of rat perseverance, reading ability, or vintage quality all three sets of data must conform to the same invariance requirements. When data fail to fit a model, Rasch measurement theory (Rasch 1960; Andrich 1988, 2010; Wright 1977, 1999) does not respond by relaxing the invariance requirements and adding, say, an item specific discrimination parameter to improve fit, as does Item Response Theory (Hambleton et al. 1991). Rather, the Rasch approach is to examine the items serving as the medium for making observations, and to change them in ways likely to produce new data conforming with theory and data model expectations.

A causal Rasch model (in which item calibrations come from theory, not data) is then doubly prescriptive (Stenner et al. 2009a, b). First, in accord with Rasch, it is prescriptive regarding the data structures that must be present:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (Rasch 1961, p. 321).

Second, causal Rasch Models (Burdick et al. 2006; Stenner et al. 2008) prescribe the values imposed by substantive theory on the item calibration estimates. Thus, the data, to be useful in making measures, must conform to both Rasch model invariance requirements *and* to substantive theory invariance requirements as specified by the theoretical item calibrations.

When data meet both sets of requirements then those data are useful not just for making measures of some vaguely defined construct but are useful for making measures of that precise construct specified by the equation that produced the theoretical item calibrations. We emphasize that these dual invariance requirements come into stark relief in the extreme case of no connectivity across stimuli or examinees (i:p). How, for example, are two readers to be measured on the same scale if they share no common text passages or items? If you read a Hunger Games novel and answer machine generated questions about it, and I read a Lord of the Rings novel and answer machine generated questions about it, how would it be possible to realize an invariant comparison of our reading abilities except by means of predictive theory? How else would it be possible to know that you read 250L better than I, and, furthermore, that you comprehended 95 % of what you read, whereas I comprehended 75 % of what I read? Most importantly, by what other means than theory would it ever be possible to reproduce this result to within a small range of error using another two completely different books as the basis of comparison?

Given that seemingly nothing is in common between the above two reading experiences, invariant comparisons might be thought impossible. Yet in the thermometer example, it is in fact a routine everyday experience for different instruments to be interpreted as informing comparable measures of temperature. Why are we so quick to accept that you have a 104 °F high grade fever and I have a 100 °F low grade fever (based on measurements from two different thermometers) and yet find the book reading example inexplicable? Is it because there are fundamental differences between physical science measurement and behavioral science measurement? No! The answer lies in well-developed construct theory, rigorously established instrument engineering principles, and uniform metrological conventions (Fisher 2009).

Clearly, each of us has had ample confirmation that weight denominated in pounds and kilograms can be well measured by any reputable manufacturer's bathroom scale. Experience with diverse bathroom scales has convinced us that, within a pound or two of error, these instruments will produce not just invariant relative differences between two persons but will also meet the more stringent expectation of invariant absolute magnitudes for each individual independent of instrument. Over centuries, instrument engineering has steadily improved to the point that for most purposes "uncertainty of measurement" (usually interpreted as the standard deviation of a distribution of imagined or actual replications taken on a single person) can be effectively ignored for most bathroom scale applications. And, quite importantly, by convention (i.e., the written or unwritten practice of a community) weight is denominated in standardized units (kilograms or pounds). The choice of any given unit is arbitrary, but what is decisive is that a unit is agreed to by

the community and is slavishly maintained through consistent implementation, instrument manufacture, and reporting. At present, language ability (reading, writing, speaking, and listening) does not enjoy a common construct definition, nor a widely promulgated set of instrument specifications, nor a conventionally accepted unit of measurement. The challenges that must be addressed in defining constructs, specifying instrument characteristics, and standardizing units include cultural assumptions about number and objectivity, political challenges in shaping legislation, resource allocation, and the expectations and procedures of social scientists (Fisher 2012, n.d.). In this context, the Lexile Framework for Reading (Stenner et al. 2006) stands as an exemplar of how psychosocial measurement can be unified in a manner precisely parallel to the way unification was achieved for length, temperature, weight and dozens of other useful attributes (Stenner and Stone 2010).

A causal (constrained) Rasch model (Stenner et al. 2009a, b) that fuses a substantive theory to a set of axioms for conjoint additive measurement affords a much richer context for the identification and interpretation of anomalies than does a descriptive i.e. unconstrained Rasch model. First, with the measurement model and the substantive theory fixed, anomalies are understood as problems with the data. Attending to the data ideally leads to improved observation models (e.g. new task types) that reduce unintended dependencies and variability. An example of this kind of improvement in measurement was realized when the Duke of Tuscany put a top on some of the early thermometers, thus reducing the contaminating influences of barometric pressure on the measurement of temperature. In contrast with the descriptive paradigm dominating much of education science, the Duke did not propose parameterizing barometric pressure in the model in the hope that the boiling point of water at sea level, as measured by open top thermoscopes, would then match the model expectations at 3000 ft above sea level (for more on the history of temperature see Chang 2004).

Second, with both model and construct theory fixed our task is to produce measurement outcomes that fit the invariance requirements of both measurement theory and construct theory. By analogy, not all fluids are ideal as thermometric fluids. Water, for example, is non-monotonic in its expansion with increasing temperature. Mercury, in contrast, has many useful properties as a thermometric fluid. Does the discovery that not all fluids are useful thermometric fluids invalidate the concept of temperature? No! In fact, a single fluid with the necessary properties would suffice to validate temperature as a useful construct. The existence of a persistent invariant framework makes it possible to identify anomalous behavior (water's strange behavior) and interpret it in an expanded theoretical framework (Chang 2004).

Analogously, finding that not all reading item types produce data that conform to the dual invariance requirements of a Rasch model and the Lexile theory does not invalidate either the axioms of conjoint measurement theory or the Lexile reading theory. Rather, the anomalous behaviors of some kinds of text (recipes, and, poems) are open invitations to expand the theory to account for these deviations from expectation. Notice here the subtle shift in perspective. We do not need to find 1000 unicorns; one will do to establish the reality of the class. The finding that reader

behavior on a minimum of two types of reading tasks can be regularized by the joint actions of the Lexile theory and a Rasch model is sufficient evidence for the existence of the reading construct (Markus and Borsboom 2013). Of course, actualizing this scientific reality to make the reading construct a universally uniform and available object in the world requires the investment of significant social, legal, and economic resources (Fisher 2005, 2009, 2000a, b, 2011, n.d.; Fisher and Stenner n.d.).

Equation (1) is a causal Rasch model for dichotomous data, which sets a measurement outcome (expected score) equal to a sum of modeled probabilities

$$\text{Expected score} =: \sum \frac{e^{(b-d_i)}}{1+e^{(b-d_i)}} \tag{1}$$

The measurement outcome is the dependent variable and the measure (e.g., person parameter, b) and instrument (e.g., the parameters d_i pertaining to the difficulty d of item i) are independent variables. The measurement outcome (e.g., count correct on a reading test) is observed, whereas the measure and instrument calibrations are not observed but can be estimated from the response data and substantive theory, respectively. When an interpretation invoking a predictive mechanism is imposed on the equation, the right-side variables are presumed to characterize the process that generates the measurement outcome on the left side. The symbol=: was proposed by Euler circa 1734 to distinguish an algebraic identity from a causal identity (right hand side causes the left hand side). This symbol (=:) was reintroduced by Judea Pearl and can be read as indicating that manipulation of the right hand side via experimental intervention will cause the prescribed change in the left hand side of the equation. Simple use of an equality (=) does not signal a causal interpretation of the equation.

A Rasch model combined with a substantive theory embodied in a specification equation provides a more or less complete explanation of how a measurement instrument works (Stenner et al. 2009a, b). A Rasch model in the absence of a specified measurement mechanism is merely a probability model. A probability model absent a theory may be useful for describing or summarizing a body of data, and for predicting the left side of the equation from the right side, but a Rasch model in which instrument calibrations come from a substantive theory that specifies how the instrument works is a causal model. That is, it enables prediction after intervention.

Below we summarize two key distinguishing features of causal Rasch models and highlight how these features can contribute to improved ENL, EFL and ESL measurement.

1. First, causal Rasch models are individually centered, meaning that a person's measure is estimated without recourse to any data on other individuals. The measurement mechanism that transmits variation in the language attribute (within person over time) to the measurement outcome (count correct on a reading test) is hypothesized to function the same way for every person. This hypothesis is testable at the individual level using Rasch Model fit statistics.

2. Figuring prominently in the measurement mechanism for language measurement is text complexity. The specification equation used to measure text complexity is hypothesized to function the same way for most text genres and for readers who are ENL, EFL and ESL. This hypothesis is, also, testable at the individual level but aggregations can be made to examine invariance over text types and reader characteristics.

EdSphere™ Reader App

The data for computing empirical text complexity measures came from the reader appliance in EdSphere™. Students access tens of millions of professionally authored digital text by opening EdSphere™ and clicking on the Reader App. Digital articles are drawn from hundreds of periodicals including Highlights for Children, Boys Life, Girls Life, Sports Illustrated, Newsweek, Discovery, Science, The Economist, Scientific American, etc. Such a large repository of high quality informational text is required to immerse students with widely varying reading abilities in daily deliberate practice across the K-16 education experience.

Students use three search strategies to locate articles targeted at their Lexile level: (1) click on suggested topics, (2) click the icon "Surprise Me", or the most frequently used method (3) type search terms into "Find a Book or Article" (see Fig. 1). In the example below, a 1069L reader typed "climate change" in the search box and found 13,304 articles close to her reading level. The first article is an 1100L 4-pager from *Scientific American* with a short abstract.

Readers browse the abstracts and refine the search terms until they find an appropriate length article about their interest topic (or a teacher assigned topic) at their reading level. Within one second of selecting an article, the machine builds a set of embedded semantic cloze items. Students choose from the four options that appear at the bottom of the page. The incorrect options have similar difficulty and part of speech to the correct answer. The answer is auto-scored and the correct answer is immediately restored in the text and color coded as to whether the student answered correctly or incorrectly.

Three instructional supports are built into the Reader App to facilitate comprehension. *First*, suggested strategies are presented to students during the reading process. *Second*, students have access to an in-line dictionary and thesaurus (one click access). Finally, a text-to-speech engine has been integrated into EdSphere, allowing words, phrases or sentences to be machine spoken to the reader.

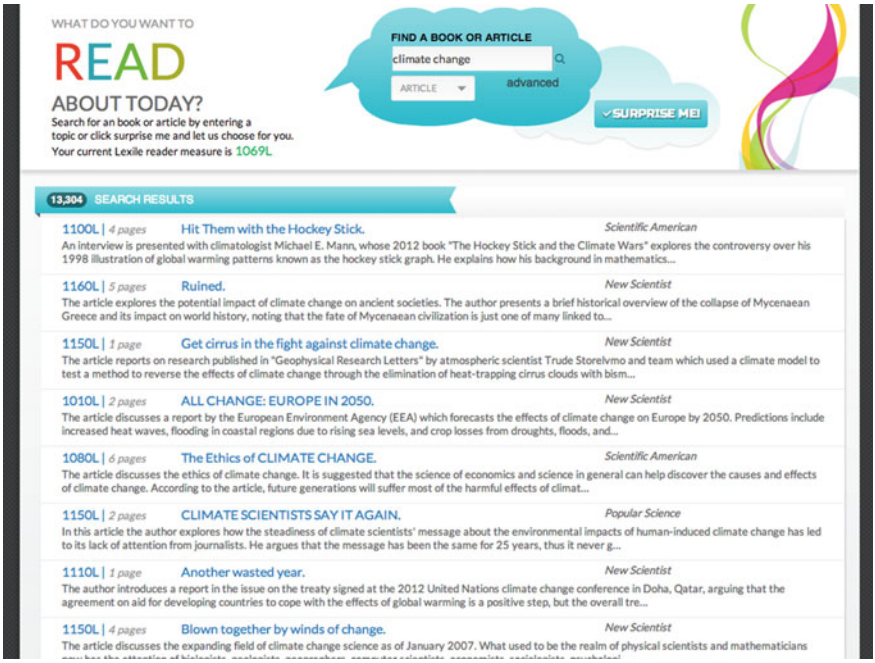


Fig. 1 Results from student's keyword search for articles about climate change; results include, publication titles, publication dates, and/or page length

Text Complexity 719's with Artifact Correction

Figure 2 presents the results of a multiyear study of the relationship between theoretical text complexity as measured by the Lexile Analyzer (freely available for non-commercial use at Lexile.com) and empirical text complexity as measured by the Edpshere™ platform. Each of the 719 articles included in this study was evaluated by the analyzer for semantic demand (log transformed frequency of each word's appearance in a multibillion word corpus) and syntactic demand (log transformed mean sentence length). The text preprocessing, what constitutes a *word*, involves thousands of lines of code. Modern computing enables the measurement of the Bible or Koran in a couple of seconds.

The Edpshere™ platform enables students to select articles of their choosing from a collection of over 100 million articles which have been published and measured over the past 20 years. As a student's reading ability grows a 200L window moves up the scale (100L below the student's ability to 100L above) and all articles relevant to a reader's search term that have text complexity measures in the window are returned to the reader. The machine generates a four choice cloze every 70–80 words and the count correct combined with the readers Lexile measure

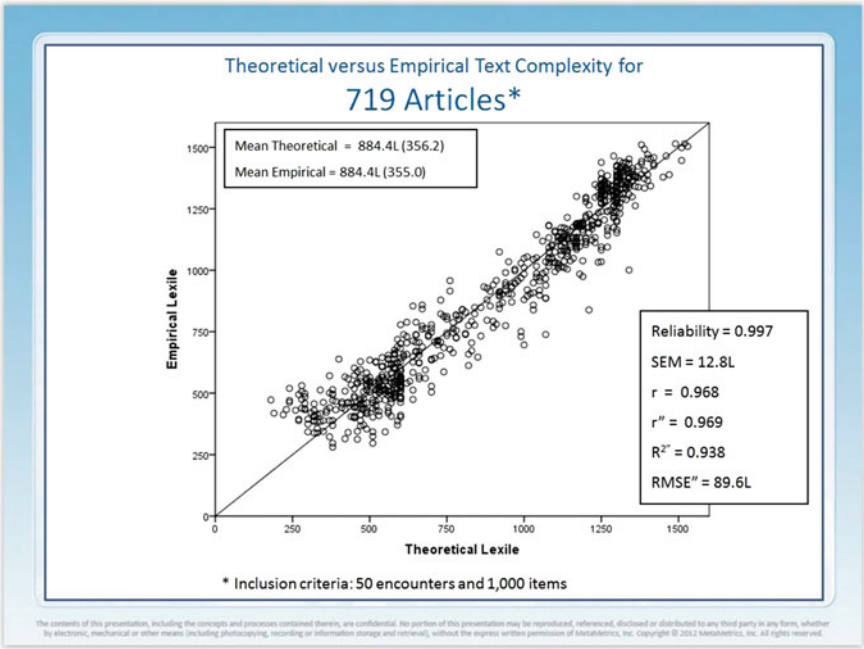


Fig. 2 Plot of Theoretical and Empirical text complexity measures

is used to compute an empirical text complexity for the article averaged over at least 50 readers and at least 1000 items.

The 719 articles chosen for this study were the first articles to meet the dual requirements of at least 50 readers and at least 1000 item responses. Well estimated reader measures were available prior to the encounter between an article and a reader. Thus, each of the articles has a theoretical text complexity measure from the Lexile Analyzer and an empirical text complexity from EdSphere. The correlation between theory and empirical text complexity is r = 0.968 (r^2 = 0.938).

Connecting Causal Rasch Models to theories of language development (Hanlon 2013; Swartz et al. 2015) has made extensive use of Ericsson's theory of deliberate practice in the acquisition of language expertise (Ericsson 1996, 2002, 2006). Deliberate practice is a core tenant of Ericsson's theory of expertise development. Hanlon (2013) distills five core principles of deliberate practice in the development of reading ability: targeted practice reading text that is not too easy and not too hard, (2) real time corrective feedback on embedded response requirements, (3) distributed practice over a long period of time (years, decades), (4) intensive practice that avoids burnout and (5) self-directed options when one on one coaching is not available. Each of these principles, when embedded into instructional technologies, benefits from individually centered psychometric models in which, for example, readers and text are measured in a common unit.

Swartz et al. (2015) provide a complete description of EdSphere, its history and components. The EdSphere Technology is designed to immerse students in deliberate practice in reading, writing, content vocabulary, and practice with conventions of standard English: "These principles of deliberate practice are strengthened by embedding psychometrically sound assessment approaches into learning activities. For example students respond to cloze items while reading, compose short and long constructed responses in response to prompts, correct different kinds of convention errors (i.e. spelling, grammar, punctuation, capitalization) in authentic text, and select words with common meanings from a Thesaurus-based activity. Each item encountered by students is auto-generated and auto-scored by software. The results of these learning embedded assessments are especially beneficial when assessment item types are linked to a developmental scale" (Swartz et al. 2015).

Figure 3 is an individual-centered reading growth trajectory denominated in Lexiles. All data comes from EdSphere. Student 1528 is an ESL seventh grade male (first language Spanish) who read 347 articles of his choosing (138,695 words) between May 2007 and April 2011. Each solid dot corresponds to a monthly average Lexile measure. The growth curve fits the monthly means quite well, and this young man is forecasted (big dot on the far right of the figure) to be a college-ready reader when he graduates from high school in 2016. The open dots distributed around 0 on the horizontal axis are the expected performance minus observed performance (in percents) for each month. Expected performance is computed using the Rasch model and inputs for each article's text complexity and

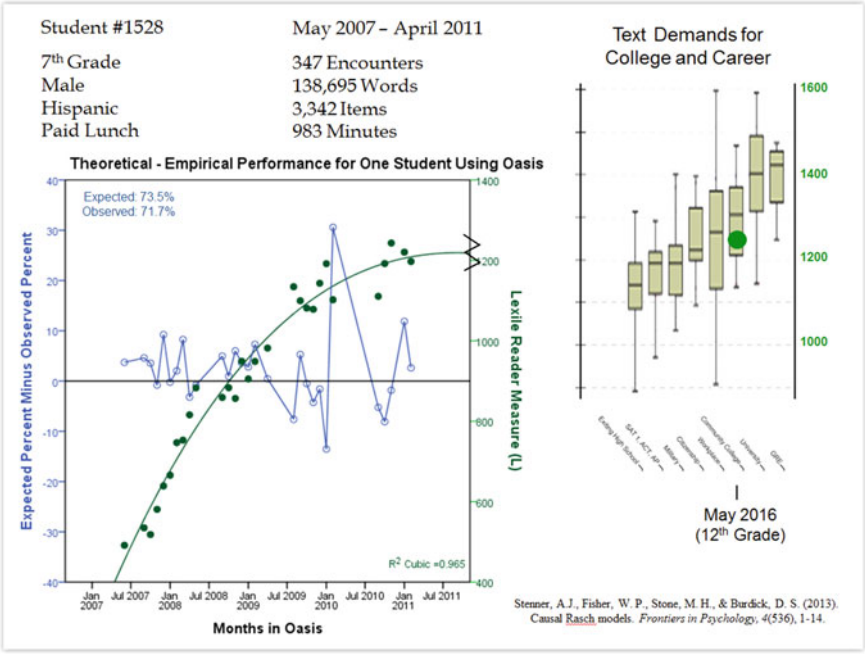


Fig. 3 An individual-centered reading growth trajectory denominated in Lexiles

the updated readers ability measure. Given these inputs, EdSphere forecasts a percent correct for each article encounter. The observed performance is the observed percentage correct for the month. The difference between what the substantive theory (Lexile Reading Framework) in cooperation with the Rasch model expects and what is actually observed is plotted by month. The upper left hand corner of the graphic summarizes the expected percentage correct over the four years (73.5 %) and observed percentage correct (71.7 %) across the 3342 items taken by this reader. Note that EdSphere is dynamically matching text complexity of the articles the reader can choose to the increasing reader ability over time. So, this graphic describes a within-person (intra-individual) test of the quantitative hypothesis: Can EdSphere trade-off a change in reader ability for a change in text complexity to hold constant the success rate (comprehension)? For this reader, the answer appears to be a resounding yes! This trade-off or cancellation affords an intra-personal test of the quantitative hypothesis (Michell 1999).

Figure 4 is a graphical depiction of the 99 % confidence interval for the artifact corrected correlation between theoretical and empirical text complexity. The artifacts included measurement error, double range restriction and construct invalidity. The artifact corrected correlation (coefficient of theoretical equivalence) is slightly higher than r = 1.0 suggesting that the Lexile Theory accounts for all of the true score variation in the empirical text complexity measures. The reader may be puzzled about how a correlation can be higher than r = 1.0, of course it can't be, but

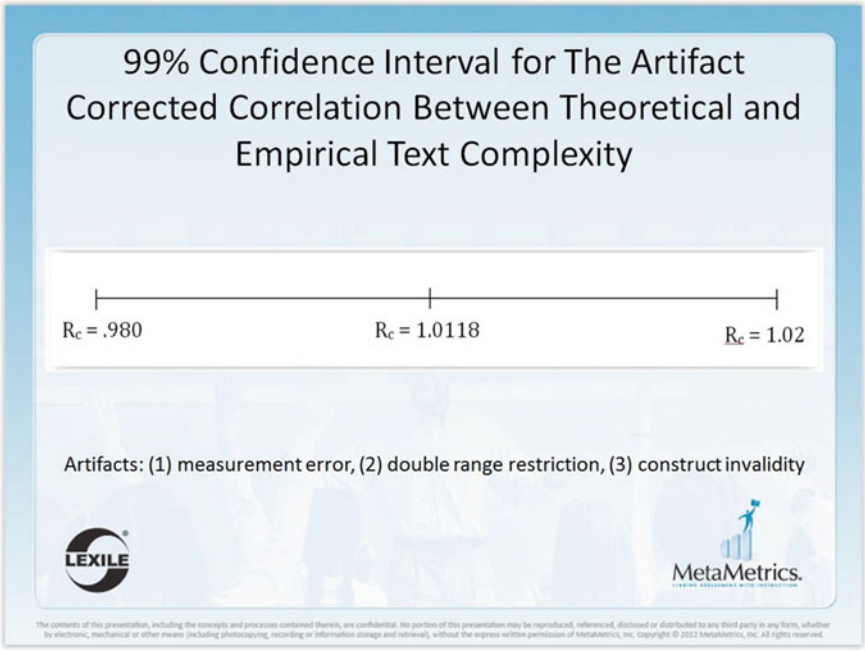


Fig. 4 Artifact corrected correlation between theory observed text complexity

an artifact corrected correlation can be if one or more artifactors used in the process are, perhaps due to a sampling error, lower than their population values.

In the temperature example, a uniform increase or decrease is the amount of soluble additive in each cavity, changes the correspondence table that links the number of cavities that turn black to degrees Fahrenheit or Celsius. Similarly, an increase or decrease in the text demand (Lexile) of the passages used to build reading tests, predictably alters the correspondence table that links count correct to Lexile reader measure. In the former case, a temperature theory that works in cooperation with a Guttman model produces temperature measures. In the latter case, a reading theory that works in cooperation with a Rasch model produces reader measures. In both cases, the measurement mechanism is well understood, and we exploit this understanding to address a vast array of counterfactuals (Woodward 2003). If things had been different (with the instrument or object of measurement), we could still answer the question as to what then would have happened to what we observe (i.e., the measurement outcome). It is this kind of relation that illustrates the meaning of the expression, "There is nothing so practical as a good theory" (Lewin 1951).

Notes

1. The NexTemp® thermometer is a small plastic strip pocked with multiple enclosed cavities. In the Fahrenheit version, 45 cavities arranged in a double matrix serve as the functioning end of the unit. Spaced at 0.2 °F intervals, the cavities cover a range from 96.0 °F to 104.8 °F. Each cavity contains three cholesteric liquid crystal compounds and a soluble additive. Together, this chemical composition provides discrete and repeatable change-of-state temperatures consistent with the device's numeric indicators. Change of state is displayed optically (cavities turn from green to black) and is easily read.
2. Text complexity is predicted from a construct specification equation incorporating sentence length and word frequency components. The squared correlation of observed and predicted item calibrations across hundreds of tests and millions of students over the last 15 years averages about 0.93. Recently available technology for measuring reading ability employs computer-generated items built "on-the-fly" for any continuous prose text in a manner similar to that described for mathematics items by Bejar et al. (2003). Counts correct are converted into Lexile measures via a Rasch model estimation algorithm employing theory-based calibrations. The Lexile measure of the target text and the expected spread of the cloze items are given by theory and associated equations. Differences between two readers' measures can be traded off for a difference in Lexile text measures to hold comprehension rate constant. When the item generation protocol is uniformly applied, the only active ingredient in the measurement mechanism is the choice of text complexity (choosing a 500L article on panda bears) and the cloze protocol implemented by the machine.

References

Andrich, D. (1988) *Rasch models for measurement* (Vols. 07-068). Sage University Paper Series on Quantitative Applications in the Social Sciences, Beverly Hills, California: Sage Publications.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*, 1–16.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika, 75*, 292–308.

Bejar, I., Lawless, R., Morley, M., Wagner, M., Bennett, R., & Revuelta, J. A. (2003). feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment 2*(2003), 1–29. http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1663.

Burdick, D., Stone, M., & Stenner, A. J. (2006). The combined gas law and a Rasch reading law. *Rasch Measurement Transactions, 20*, 1059–1060.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. New York: Oxford University Press.

Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games* (pp. 1–50). Mahwah, NJ: Erlbaum.

Ericsson, K. A. (2002). Attaining excellence through deliberate practice: Insights from the study of expert performance. In M. Ferrari (Ed.), *The pursuit of excellence in education* (pp. 21–55). Hillsdale, NJ: Erlbaum.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 683–703). Cambridge, UK: Cambridge University Press.

Fisher, W., Jr. (1997). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement, 1*, 87–113.

Fisher, W., Jr. (1999). Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. *Journal of the Louisiana State Medical Society, 151*, 566–578.

Fisher, W., Jr. (2000a). Rasch measurement as the definition of scientific agency. *Rasch Measurement Transactions, 14*, 761.

Fisher, W., Jr. (2000b). Objectivity in psychosocial measurement: What, why, how. *Journal of Outcome Measurement, 4*, 527–563.

Fisher, W., Jr. (2005). Daredevil barnstorming to the tipping point: New aspirations for the human sciences. *Journal of Applied Measurement, 6*, 173–179.

Fisher, W., Jr. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement, 42*, 1278–1287.

Fisher W., Jr. (2011). Bringing human, social, and natural capital to life: Practical consequences and opportunities. In N. Brown, B. Duckor, K. Draney, & M. Wilson (Eds.), *Advances in Rasch Measurement* (Vol. 2, pp. 1–27). Maple Grove, Minnesota: JAM Press.

Fisher W., Jr. (2012). What the world needs now: A bold plan for new standards, *Standards Engineering 64*, in press.

Fisher W., Jr. NIST critical national need idea White Paper: Metrological infrastructure for human, social, and natural capital, Retrieved 6 March 2012 from http://www.nist.gov/tip/wp/pswp/upload/202_metrological_infrastructure_for_human_social_natural.pdf. Washington, DC: National Institute for Standards and Technology.

Fisher W., Jr., & Stenner, A. J. Metrology for the social, behavioral, and economic sciences (Social, Behavioral, and Economic Sciences White Paper Series). Retrieved 6 March 2012, from http://www.nsf.gov/sbe/sbe_2020/submission_detail.cfm?upld_id=36, Washington, DC: National Science Foundation.

Fisher, W., Jr., Harvey, R., & Kilgore, K. (1995). New developments in functional assessment: Probabilistic models for gold standards. *NeuroRehabilitation, 5*, 3–25.

Hambleton, R., Swaminathan, H., & Rogers, L. (1991). *Fundamentals of item response theory*. Newbury Park, California: Sage Publications.

Hanlon, S. T. (2013). *The relationship between deliberate practice and reading ability* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses databases (AAT 3562741).

Hunter, J. (1980). The national system of scientific measurement. *Science, 210*, 869–874.

Kuhn, T. S. (1961). The Function of Measurement in Modern Physical Science. *Isis, 52*, 161–193.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. New York: Cambridge University Press.

Lattanzio, S., Burdick, D., & Stenner, A. J. (2012). *The Ensemble Rasch Model*. Durham, NC: MetaMetrics Paper Series.

Lewin, K. (1951). *Field theory in social science: Selected theoretical papers*. New York: Harper & Row.

Markus, K. A. & Borsboom, D. (2013). *Frontiers of Test Validity Theory*. Routledge.

Michell, J. (1999). *Measurement in Psychology*. Cambridge University Press.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests (Reprint, with Foreword and Afterword by B. Wright, University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV* (pp. 321–334). Berkeley, California: University of California Press.

Stenner, A. J., & Stone, M. (2010). Generally objective measurement of human temperature and reading ability: Some corollaries. *Journal of Applied Measurement, 11*, 244–252.

Stenner, A. J., Burdick, H., Sanford, E., & Burdick, D. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement, 7*, 307–322.

Stenner, A. J., Burdick, D., & Stone, M. (2008). Formative and reflective models: Can a Rasch analysis tell the difference? *Rasch Measurement Transactions, 22*, 1152–1153.

Stenner, A. J., Stone, M., & Burdick, D. (2009a). The concept of a measurement mechanism. *Rasch Measurement Transactions, 23*, 1204–1206.

Stenner, A. J., Stone, M., & Burdick, D. (2009b). Indexing vs. measuring. *Rasch Measurement Transactions, 22*, 1176–1177.

Stenner, A. J., Fisher, W. P., Stone, M. H. & Burdick, D. S. (2013). Causal Rasch Models. *Frontiers in Psychology, 4*, 1–14.

Swartz, C. W., Hanlon, S. T., Stenner, A. J., & Childress, E. L. (2015). An approach to design-based implementation research to inform development of EdSphere®: A brief history about the evolution of one personalized learning platform. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development*. IGI Global: Hersey, PA.

Taagepera, R. (2008). *Making social sciences more scientific: The need for predictive models*. New York: Oxford University Press.

Thurstone, L. L. (1926). The Scoring of Individual Performance. *Journal of Educational Psychology, 17*, 446–457.

Woodward, J. (2003). *Making things happen* (p. 410). Oxford: Oxford University Press. pp. vi.

Wright, B. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.

Wright, B. (1999). Fundamental measurement for psychology. In S. Embretson & S. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 65–104). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Constructing the Human Figure Drawing Continuum: One Scale is 'Good Enough'

Claire Campbell and Trevor Bond

Introduction

Historically, drawing has been considered an innate form of expression for humans (Fowlkes 1980). Prior to the development of verbal communication systems, our pre-historic ancestors used drawings to convey information. Indeed, the first crude form of written language—known as 'proto-writing'—consisted of tiny, intricate drawings rather than early forms of letters or numerals (Houston 2004). Pre-historic cave drawings and present-day young children's drawings could be considered to share a similar underpinning motive; both used drawing to convey information, understanding and knowledge during a time when other forms of communication were not yet fully mastered (Fowlkes 1980). Indeed, today's young children characteristically delight in drawing representations of what they know and understand about the world, while still acquiring verbal and written communication skills (Di Leo 1970). Whilst the notion of children's 'drawings' typically evokes discussions about art, creativity and imagination, drawings reveal much more developmental and intellectual content than most realise.

Children's drawings have been researched from a variety of perspectives for over 150 years. Children's drawings of the human figure are amongst the most researched (Cox 1992; Kellogg 1967, 1970; Koppitz 1968). This is not surprising to many, as young children's first attempts at creating a human figure drawing (HFD) —known as cephalopods (see Fig. 1)—are amongst the most distinctive and recognisable aspects of early childhood. The wealth of research indicates that there is a general developmental sequence to children's HFDs (Cox 1992, 1997; Di Leo 1970, 1973; Luquet 1913; Maley 2009; Mavers 2011; Piaget and Inhelder 1956, 1971). That is, similar to how children typically crawl, walk and then run sequentially in the domain of gross motor development (Goodway et al. 2012),

C. Campbell (✉) · T. Bond
College of Arts, Society and Education, James Cook University, Townsville, Australia
e-mail: claire.campbell2@jcu.edu.au

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_2

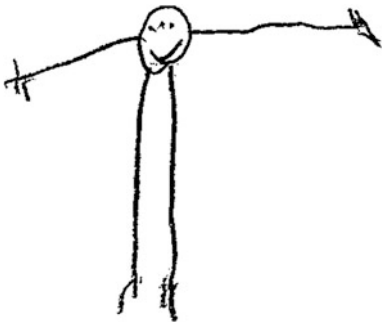


Fig. 1 Example of a cephalopod drawing by a young child

Fig. 2 Examples of children's drawings at three stages (Piaget and Inhelder 1956, 1971)

children sequentially make scribbles in the 'fortuitous realism' or 'scribbling' stage, and then 'projective' and 'imaginal representations' of humans in the 'intellectual realism' stage, followed by drawings with more accurate proportions and perspective in the 'visual realism' stage in the area of HFD development (see Fig. 2) (Piaget and Inhelder 1956, 1971).

Piagetian theory provides a useful lens for investigating young children's drawings. Jean Piaget (1896–1980) was the first to systematically study children's cognitive development and the findings of his Genevan school have had significant impact on the fields of education and psychology. Piagetian theory—which describes children's cognitive development as progressing through an invariant sequence of age-related (not age-dependent) stages of thinking: sensorimotor (birth to toddlers); preoperational (early years education); concrete operational (primary school) and formal operational (secondary school and onwards)—is well known in educational and developmental psychology circles. However, many parents, carers of young children and lay people remain unfamiliar with his theories. Consequently, when children display behaviours that could be considered characteristic of the preoperational stage of thought—such drawing of cephalopods—many people dismiss this as just some of the 'bizarre' things about early childhood. Indeed, many

parents and carers might attempt to teach young children to include necks, torsos, shoulders and similar features in their HFDs or, in a related example, to explain to them that they do not have more cake to eat now that it has been cut into pieces as opposed to its being left whole. However, adherents to Piagetian theory would suggest that these efforts would be to no avail; until the child develops more organised mental structures, the benefit of this sort of didactic teaching cannot be realised (Piaget 1971).

Piaget and Inhelder (1956, 1971) examined young children's 'representations', or drawings, as an extension of their earlier work on cognitive development. In brief, their theory suggests that the 'borderlines' between the child's perception of an object, the child's mental representation of that object and the child's actual pictorial representation of that object create *discontinuities* in the drawing process. As children's thinking progress through the developmental stages, become increasingly more able to transition with *continuity* through these borderlines and, consequently, produce more comprehensive and realistic pictorial representations (Piaget and Inhelder 1956, 1971). As young children's thinking transitions toward the concrete operational stage of thought, their perception becomes progressively decentred; they are more able to integrate information from a variety of sources and reorganise their knowledge into more sophisticated systems. As a result, older and more experienced children become increasingly able to see others' points of view, consider multiple aspects of a problem at one time and produce more logical and comprehensive explanations and drawings.

Many researchers agree on the links between young children's HFDs and their levels of cognitive development and understanding (Anning and Ring 2004; Cox 1992; Di Leo 1973; Goodenough 1926; Harris 1963; Piaget and Inhelder 1956, 1971). The most influential work in this body of knowledge is held to be that originally undertaken by Goodenough (1926) and Harris (1963).

Background

Development of the Goodenough Draw-a-Man Test (GDAMT)

Florence Goodenough (1886–1959) was an American teacher interested in intelligence and conventional intelligence (IQ) tests (Goodenough 1949; Goodenough and Tyler 1959). She worked under the supervision of Lewis Terman at Stanford University developing the Stanford-Binet Intelligence Quotient test for children (Thompson 1990). Her doctoral research (Goodenough 1924) into young children's human figure drawings resulted in the GDAMT, the world's first non-verbal assessment for inferring children's levels of cognitive development and understanding via the details included in their drawings of an adult male human figure. Her published doctoral thesis, *Measurement of Intelligence by Drawings*

(Goodenough 1926), detailed the development of the 51-item test and explained the strategic selection of an adult 'male human' as subject matter. Goodenough's research indicated that other drawings, including those of women, houses, animals and the like, were not sufficiently 'uniform' and thereby produced erratic results using the statistical methods of the time (Goodenough 1926). Unlike the other non-verbal assessments available at that time, the GDAMT exhibited high reliability and validity, was easy to administer, and correlated well with other intelligence tests (Goodenough 1926; Harris 1963).

Development of the Goodenough-Harris Drawing Test (GHDT)

Goodenough's doctoral student, Dale Harris (1915–2007), revised and extended the original GDAMT; it is now known as the GHDT. Harris made two key changes: (1) he extended the original 51-item GDAMT to a 73-item Draw-a-Man (DAM) sub-test; and (2) he included a 71-item Draw-a-Woman (DAW) sub-test and a Self-Portrait sub-test (73-item SPM sub-test for male children and 71-item SPF sub-test for female children). As with the original GDAMT, the GHDT yielded high reliability and validity and correlated well with the other more conventional intelligence tests using the statistical methods available at the time (Harris 1963). Notwithstanding, Harris added the DAW and SPM/SPF sub-tests even in the face of a complete lack of empirical evidence indicating that a single drawing of a man was actually insufficient for the assessment task. Furthermore, Harris did not verify empirically the effectiveness of the tripled data collection load (i.e., the collection of three drawings instead of one) for each child.

Research Objective

This study adopted the Rasch model for measurement (Rasch 1960) to (1) examine the psychometric properties of the GHDT and young children's HFDs, and (2) investigate what additional information might be revealed by Harris's additional items for the DAM sub-test, as well as the DAW and SP sub-tests.

Method

Sample

A sample of 107 children of different ages and abilities (aged between 4 and 10 years) were recruited from a Preparatory to Year 12 school in Queensland,

Australia (Preparatory, or 'Prep', is the year level prior to the first year of formal education). A total of 738 HFDs were collected from these children over three phases of data collection, approximately six months apart. The sample included children from a range of socioeconomic backgrounds, however, the ethnic backgrounds of the participants were not as diverse. Most of the sample was of Caucasian descent with approximately 3 % of children from Australian Aboriginal, Torres Strait Islander and Asian backgrounds. The children's HFDs were not examined according to socioeconomic status or ethnicity. All participants had informed parental consent to participate in the study.

Data Collection

The children's HFDs were collected via the published GHDT administration procedures, which simply require children (either individually or in small groups) to make three drawings, one each of a man, a woman and a self-portrait, to the best of their ability. Provided that children attempt to draw a whole person (that is, not a 'bust'-type portrait), there is no 'right' or 'wrong' type of drawing. During the small group data collection procedures, children were seated so that they could not 'copy' from another's drawings. The GHDT administration does not have a time limit, although most children complete all three drawings in approximately 10–15 minutes. All HFDs were examined and scored (by the first author) in alignment with the two GHDT scoring guides: a 73-item guide for the DAM sub-test and a 71-item guide for the DAW sub-test. Whilst the scoring guides share 50 common items, most item numbers do not correspond across the guides. Moreover, the SPM and SPF sub-tests do not have dedicated scoring guides; they are scored merely using the DAM or DAW scoring guide as appropriate to the reported gender of the child.

Data Analysis

The qualitative drawings were converted into quantitative data so that the Rasch model could be applied. This was quite a straightforward process as the scoring guides used a dichotomous (0, 1) system for each drawing criterion. That is, if the item was judged as absent from the HFD or present but not in fulfilment of the criterion, zero (0) credit was assigned to that item number. Conversely, if the item was present in the HFD and satisfying the criterion, one (1) credit was assigned to that item number. Once scoring was complete, data files were created using Excel® software and then submitted to WINSTEPS® version 3.68.0 (Linacre 2009) for Rasch analysis.

Results

Overview

Rasch analysis of the data produced a range of output including variable maps and summary statistics. The results of the Rasch analysis of the DAM, DAW, SPM and SPF sub-tests revealed quite remarkable similarity. The equal-interval logit measurement scales produced by the Rasch analysis of the DAM and DAW sub-tests each spanned a range of 14 logits (from −8 to +6 logits). The SPM and SPF logit scales were only two logits shorter spanning from −7 to +5. Interestingly, most of the 50 test items common to all four sub-tests (e.g., head, eyes, trunk/body, legs, nose 2D) were located in similar positions along the logit scales across all four variables maps for the DAM, DAW, SPM and SPF sub-tests.

A selection of key summary statistics produced by the Rasch analyses of the GHDT sub-test data is presented together in Table 1. The total number of items and persons, means, standard deviations, reliability, separation and number of underfitting and overfitting items and persons are displayed for the each of the four GHDT sub-tests, respectively.

Unidimensionality

The comprehensive work completed by Goodenough (1926) and Harris (1963) was completed using the analytical methods appropriate for the time. Subsequently,

Table 1 Selected Rasch analyses summary statistics for the four GHDT sub-tests

	DAM	DAW	SPM	SPF
Items				
N	73	71	73	71
Mean	0.00	0.00	0.00	0.00
SD	2.90	2.77	2.48	2.71
R	0.98	0.99	0.97	0.98
Separation	7.89	8.37	5.48	6.72
Underfitting (n)	10	6	5	12
Overfitting (n)	9	12	2	7
Persons				
N	246	246	99	147
Mean	−1.26	−1.18	−1.22	−0.88
SD	1.45	1.43	1.3	1.41
R	0.93	0.92	0.89	0.93
Separation	3.55	3.50	2.89	3.54
Underfitting (n)	10	6	5	4
Overfitting (n)	2	4	0	5

neither young children's HFDs, nor the GHDT, have been investigated comprehensively from a modern test theory perspective. Correspondingly, thus far it has presumed that all items in the GHDT sub-tests contribute meaningfully to the investigation of a single underlying drawing construct. The Rasch model's unidimensionality principle requires that all items comprising the GHDT investigate only one construct at a time—in this case, the development of children's HFDs. The Rasch model's fit statistics and other diagnostic indicators helped to determine that most items and persons under investigation satisfied this unidimensionality requirement. Rasch analyses of the data indicated only a small number of erratic (i.e., underfitting) items for each of the GHDT sub-tests (see Table 1). Similarly, even fewer items were detected as being overly predictable (i.e., overfitting) across the four sub-tests (see Table 1). The person fit statistics revealed a very small number of 'misfitting' person measures across each of the GHDT sub-tests; however, there were no children who consistently presented misfitting performances.

GHDT Sub-test Comparisons

Given that the GHDT did not breach the Rasch model's unidimensionality principle, the second step in this study was to investigate what additional information, if any, was revealed by Harris's additional DAW and SPM/SPF sub-tests. Closer inspection of the person fit statistics and ability estimates revealed that children performed almost identically on all three GHDT sub-tests that they completed. Consequently, very little additional information is revealed by the DAW and the SP sub-tests over that already revealed by the DAM sub-test. Given that mean person estimates close to zero (0) are indicative of a well-matched test (Bond and Fox 2015), the SPM (0.42) and SPF (0.38) sub-tests were found to be better targeted to this sample of children than were the DAM (−1.26) and DAW (−1.18) sub-tests. Despite this, children tended to receive slightly lower raw scores for their SP drawings than they did for their drawings of adult males and females. This could be linked to the fact that the SP sub-tests are scored using the applicable adult 73-item DAM or 71-item DAW scoring guides. These guides were designed for scoring drawings of adults and contain item-scoring criteria that should be considered problematic for young children's drawings of themselves. In particular, the DAW scoring guide contains a number of items pertaining to jewellery, clothing, high heel shoes, breasts, hips and the like, which—understandably—might prove problematic for several developmental, cultural and societal reasons. A young girl's self-portrait might not include several (or all) of these items, as they are simply irrelevant developmentally, culturally and/or socially. However, the scoring of such a SPF drawing using the DAW scoring guide assumes that these items are 'absent' due to a lack of understanding, rather than inappropriateness due to the drawer's level of physical development, cultural circumstances and/or social background. It

was issues such as these that prompted the development of a modified GHDT that contained only items applicable to drawings of all human figures—men, women, boys and girls.

Development of the Human Figure Drawing Continuum

Eventually, there were several key factors that pointed towards the need for a more parsimonious and developmentally/culturally/socially relevant human figure drawing test: (1) the finding that similar items across all four sub-tests were plotted in similar locations along the logit scale in the person-item maps; (2) the finding that little additional information is yielded from the two additional sub-tests beyond that already revealed by a single drawing and (3) the finding that some items in the scoring guides were irrelevant or inappropriate to the scoring of drawings of children. Given that there were 50 common items across the DAM/SPM and DAW/SPF sub-test scoring guides, a key analytical task was to investigate whether all or any particular combination of those 50 (or fewer) common items could be used to measure effectively *any* HFD made by young children.

The development of the Human Figure Drawing Continuum (HFDC) commenced with an examination of the fit statistics for each of the 50 common items across the four GHDT sub-tests. The 50 common items were deemed to form a sound 'starting point' as each of them is applicable to drawings of males and females, regardless of age. After careful examination of the fit statistics, various combinations of misfitting items were removed iteratively in an effort to enhance the measurement properties of the proposed instrument, and repeated Rasch analyses were conducted. Each iteration produced output that was inspected closely for adherence to Rasch's measurement expectations until finally a particular combination of 45 items was identified as best fitting the Rasch model's requirements. That is, additional items or fewer items did not improve the person and item fit statistics, variance explained, standard deviations, person separation indices or person and item reliability values.

Common person linking, or invariance, graphs (Bond and Fox 2015) were used to verify the effectiveness of the HFDC as a scoring mechanism in comparison to each of the four GHDT sub-tests. Figure 3 displays the DAM *v.* HFDC and the SPM *v.* HFDC and Fig. 4 shows the DAW *v.* HFDC and the SPF *v.* HFDC, respectively. These graphs of the paired person measures from the 45-item HFDC against each of the 70-plus item GHDT sub-tests indicate that all comparisons yield congruent results. That is, all graphs show all plots located between the 95 % control lines indicating the invariance of the person measures within error—except for one, single, anomalous performance. This male child aged seven years was 'on the borderline' in the DAW *v.* HFDC common person linking graph (Fig. 4) with a DAW estimate of −2.66 logits (err = 0.44) and a HFDC estimate of −1.43 logits (err = 0.49). That is, he drew a 'human' more successfully than he drew a 'woman'.

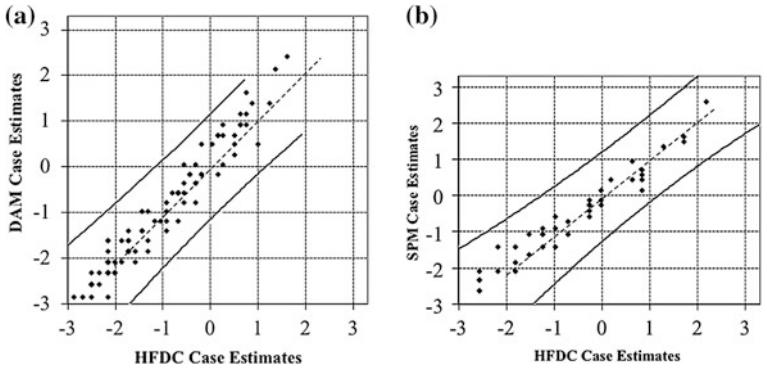


Fig. 3 Common person linking graphs: **a** 73-item DAM *v.* 45-item HFDC and **b** 73-item SPM *v.* HFDC

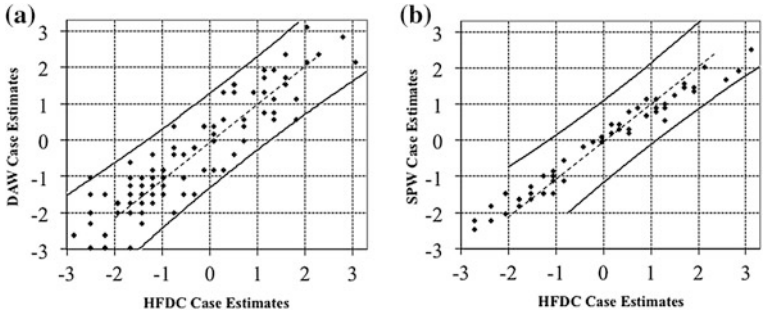


Fig. 4 Common person linking graphs: **a** 71-item DAW *v.* 45-item HFDC and **b** 71-item SPF *v.* HFDC

The graphs (Figs. 3 and 4) clearly indicate that the 45-item HFDC was almost exactly as effective as any of the 70-plus item GHDT sub-tests in assessing young children's HFDs. Indeed, it could be argued that the HFDC was the best suited and most reliable of all of the instruments. First, the HFDC yielded an item standard deviation of 3.04 indicating that it had the greatest spread of items along the logit scale in comparison to the DAM (2.90), DAW (2.77), SPM (2.48) and SPF (2.71) sub-tests. Second, the HFDC yielded an item reliability of 0.99 (i.e., congruent with the other sub-tests: DAM: 0.98; DAW: 0.99; SPM: 0.97; SPF: 0.98) suggesting that the item hierarchies are equally likely to be replicated when the instrument is administered to other suitable samples. Third, the HFDC yielded a person reliability index of 0.89, which is not meaningfully different from that produced by the DAM, DAW and SPF sub-tests (0.92) and the SPM sub-test (0.88). Fourth, the HFDC yielded the highest mean person measure (−0.59) of all the sub-tests (DAM: −1.26; DAW: −1.18; SPM: −1.22; SPF: −0.88) indicating that it is better targeted to the abilities of this sample of children. And, finally, whilst the HFDC person separation

index (2.89) was comparable with that produced by the SPM sub-test (2.71), it was slightly less than that of the DAM (3.33), DAW (3.32) and SPF (3.34) sub-tests. Despite this, the HFDC yielded over four person strata (4.17), well within the range of 'acceptable' values (Fisher 1992). That is, the HFDC can identify four measurably different performance strata of children in this sample.

Figure 5 shows the variable map produced from the Rasch analysis of the HFDC. As the focus of this analysis was on items, rather than on persons, the person identification numbers were excluded from the map. The '#' symbol on

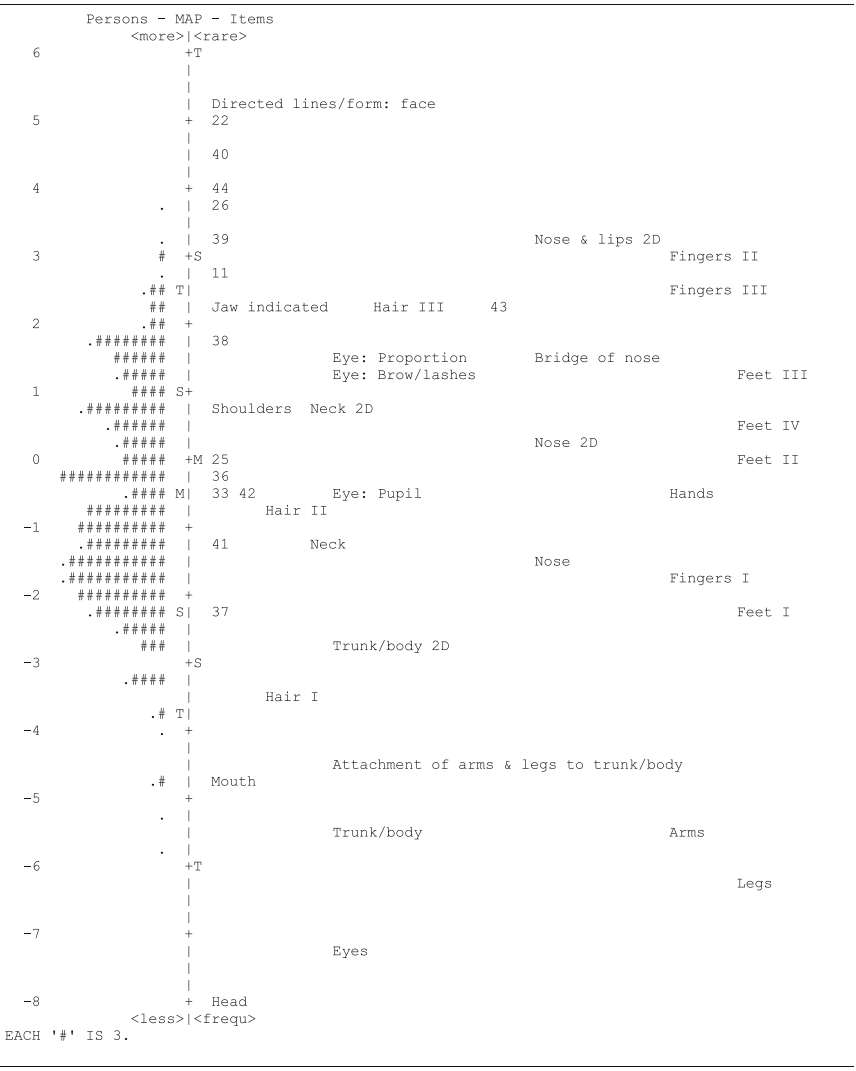


Fig. 5 Human figure drawing continuum person-item variable map

the left-hand side of the logit scale indicates that three children were located at that estimated ability level ('.' indicates $0 < n < 3$). At 14 logits, the logit scale range is just as expansive as those produced from the DAM and DAW sub-test data. Some item numbers have been replaced with the item name to help add substantive meaning to the HFDC scale.

Discussion

This study was the first to investigate the GHDT and young children's HFDs from a modern test theory perspective. The principle aims were to (1) examine the psychometric properties of the GHDT and young children's HFDs and (2) investigate the potential redundancy of Harris's additional sub-tests. Given that the GHDT and the HFDs were deemed to be psychometrically sound—and that Harris's revision and extension revealed little beyond that revealed by a single drawing—the possibility of a more parsimonious instrument was explored. Therefore, 45-item instrument HFDC was identified and verified to be just as effective as any of the four 70-plus item GHDT sub-tests.

Implications

This research has shown that, despite Harris's good intentions, children's separate drawings of men, women and themselves do not provide additional information that is useful to the test administrator. Furthermore, for this assessment, children's HFDs do not need to be scored against 70-plus criteria. A single drawing of a self-selected human figure (man, woman or child) scored against a 45-item scoring guide should reveal almost exactly the same information as three drawings and some 200-plus drawing and scoring criteria. The authors have not been able to locate information on why Harris decided to revise and extend the original GDAMT; however, it could be assumed that it was linked to societal pressure regarding inclusivity, ageism and sexism. This research has revealed that the assessment can be made more inclusive and non-sexist by merely asking children to self-select a single human figure drawing rather than by requiring children to draw one each of a man, a woman and a child.

Another implication of this research is the empirical evidence as to which HFD concepts are least and most difficult for children to include. The DAM, DAW, SPM, SPF and the HFDC all had extremely similar item hierarchies further confirming, via modern test theory, the notion that young children's HFDs are underpinned by a common developmental sequence.

Limitations

Whilst the proposed HFDC mitigates many of the limitations of the GHDT, like all research, this study has its own limitations. Foremost, the HFDC lacks the wealth of empirical support of both the GDAMT and the GHDT. Future research should involve replication to investigate whether similar results can be achieved with a larger sample of more diverse children. Also, ideally, a project could follow the development of HFDs from the very first marks produced at around 18 months of age through adulthood to investigate the HFD developmental process. Further, this research investigated only the GHDT from a modern test theory perspective. It did not correlate or co-calibrate the GHDT, or the HFDC, with other assessments (such as intelligence tests). As all data were collected from one school in Queensland, Australia, appropriate caution must be applied when transferring findings to other settings. Last, this research did not investigate the links between children's HFDs and their levels of cognitive development; it relied on the research already undertaken by Goodenough and Harris in that regard. Moreover, the data were not collected as representing the proposed HFDC, a priori; this should form a key feature of subsequent investigations. Future research could investigate young children's HFDs and their levels of cognitive development—via the administration of Piagetian conservation tasks, for example—to examine what relationship exists between the two.

Conclusion

Given that the original GDAMT was created in 1926, and Harris's revision and extension was completed in 1963, it could be concluded that the drawing assessment was due for re-investigation, particularly from a modern test theory perspective. This research confirmed that the GHDT and young children's HFDs are both apt for Rasch analysis, and that they both satisfied the Rasch model's unidimensionality principle and its other measurement requirements. Interestingly, the application of the Rasch model to the GHDT has taken us *back* to what Florence Goodenough found in the first instance—that a single human figure drawing scored according to 50 or so criteria is 'good enough'.

References

Anning, A., & Ring, K. (2004). *Making sense of children's drawings*. Berkshire, England: Open University Press.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). United Kingdom: Routledge.

Cox, M. (1992). *Children's drawings*. London: Penguin Group.

Cox, M. (1997). *Drawings of people by the under-5s*. London: Falmer Press.
Di Leo, J. H. (1970). *Young children and their drawings*. New York: Brunner/Mazel Inc.
Di Leo, J. H. (1973). *Children's drawings as diagnostic aids*. New York: Brunner/Mazel Inc.
Fowlkes, S. S. (1980). *The effect of colour on human figure drawings as related to level of social adaptability*. Unpublished Master thesis. United States of America: The Ohio State University.
Fisher, W. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions, 6*(3), 238.
Goodenough, F. L. (1924). *The intellectual factor in children's drawings*. Doctoral thesis. Stanford University.
Goodenough, F. L. (1926). *The measurement of intelligence by drawing*. New York: World Book Co.
Goodenough, F. L. (1949). *Mental testing: Its history, principles and applications*. New York: Rinehart and Company Inc.
Goodenough, F. L., & Tyler, L. E. (1959). *Developmental psychology* (3rd ed.). New York: Appleton-Century-Crofts Inc.
Goodway, J., Ozmun, J., & Galladue, D. (2012). Motor development in young children. In O. N. Saracho & B. Spodek (Eds.), *Handbook of research on the education of young children* (3rd ed.). United Kingdom: Routledge. Retrieved from http://search.credoreference.com.elibrary.jcu.edu.au/content/entry/routsmch/motor_development_in_young_children/0.
Harris, D. B. (1963). *Children's drawings as measures of intellectual maturity: A revision and extension of the Goodenough draw-a-man test*. New York: Harcourt, Brace & World Inc.
Houston, S. D. (Ed.). (2004). *The first writing: Script invention as history and process*. Cambridge: Cambridge University Press.
Kellogg, R. (1967). *The psychology of children's art*. San Diego: Random House.
Kellogg, R. (1970). *Analysing children's art*. Pal Alto, California: Mayfield.
Koppitz, E. M. (1968). *Psychological evaluation of children's human figure drawings*. New York: Grune & Stratton.
Linacre, J. M. (2009). *WINSTEPS*® Version 3.68.1 [computer software]. Chicago, IL: http://www.winsteps.com.
Luquet, G. H. (1913). *Les dessins dun enfant*. Paris: Librairie Felix Alcan.
Maley, C. (2009). *Young children's human figure drawings: An investigation using the Goodenough-Harris drawing test and the Rasch model for measurement.* Unpublished Ph.D. thesis. Australia: James Cook University.
Mavers, D. (2011). *Children's drawing and writing: The remarkable in the unremarkable*. New York: Routledge.
Piaget, J. (1971). *Biology and knowledge*. Chicago: University of Chicago Press.
Piaget, J., & Inhelder, B. (1956). *The child's conception of space*. London: Routledge and Kegan Paul.
Piaget, J., & Inhelder, B. (1971). *Mental imagery in the child*. London: Routledge & Kegan Paul Ltd.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.
Thompson, D. (1990). Florence Laura Goodenough. In A. O'Connell & N. Russo (Eds.), *Women in psychology: A bio-bibliographic sourcebook*. Westport, CT: Greenwood Press.

Using MFRM and SEM in the Validation of Analytic Rating Scales of an English Speaking Assessment

Jinsong Fan and Trevor Bond

Introduction

In second language performance assessment, both holistic and analytic rating scales are often used to award scores to test candidates. Whereas holistic scales express an overall impression of a test candidate's ability in one score, analytic scales contain a number of criteria, usually 3–5, each of which has descriptors at the different levels of the scale (Luoma 2004). Compared with holistic scales which give only one score, analytic scales have several discernible advantages including, for example, providing rich information about test candidates' language ability (e.g., Kondo-Brown 2002), and improving rating accuracy through drawing raters' attention to specific criteria of language performance (Luoma 2004). Moreover, as pointed out by Sawaki (2007), analytic scales are consistent with the current view of the multidimensional nature of language ability (see also In'nami and Koizumi 2012; Sawaki et al. 2009). As such, analytic rating scales are extensively used in L2 performance assessment such as speaking and writing (e.g., Bachman et al. 1995; Lumley 2002; Shin and Ewert 2015), particularly in the contexts where testing is more closely aligned with teaching and learning, and where rich feedback information is deemed crucial to test candidates (e.g., Sasaki and Hirose 1999).

Fulcher (1996, p. 208) argued that rating scales tend to be "*a priori* measuring instruments" in the sense that the descriptors in the rating scales are usually constructed by an expert through his or her own intuitive judgment concerning the

J. Fan (✉)
Fudan University, Shanghai, People's Republic of China
e-mail: jinsongfan@fudan.edu.cn

J. Fan
The University of Melbourne, Melbourne, Australia

T. Bond
James Cook University, Townsville, Australia

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_3

nature of language proficiency, or sometimes in consultation with a team of other language experts. Such an approach to scale development, as Fulcher (1996) continued to argue, inevitably leads to the lack of empirical underpinning. Therefore, after a rating scale has been constructed, *post hoc* validity studies are essential to verify that the descriptors are meaningful indicators of test candidates' proficiency in a specific language modality (see also Upshur and Turner 1995). This view resonates with that of Knoch (2011, p. 81) who argued that rating scales act as "the *de facto* test construct" in performance assessment. It follows therefore that construct validation of the rating scale is crucial to the establishment of the construct validity of a particular assessment. In response to this call for *post hoc* validity research of rating scales, an array of validation studies have been reported, most of which are in the domain of L2 writing assessment (e.g., Lumley 2002; Sasaki and Hirose 1999; Shin and Ewert 2015; Upshur and Turner 1999) with few focused on the assessment of L2 speaking ability (e.g., Sato 2012; Sawaki 2007; Upshur and Turner 1999).

A review of the existent research reveals that most studies have adopted either the Generalizability-theory (G-theory) which represents an extension of the Classical Test Theory (CTT) (e.g., Sato 2012; Shin and Ewert 2015) or the Many-Facets Rasch Model (MFRM), one of the Item Response Theory (IRT) models (e.g., Upshur and Turner 1999); few studies, however, have adopted a combination of two different yet complementary data analytic approaches. One exception is that of Lynch and McNamara (1998) who employed G-theory and MFRM in the development of a L2 speaking assessment for intending immigrants. As articulated by the two researchers, the G-theory is able to take all the various facets of a measurement procedure into account, and to differentiate their effects, via the estimated variance components, on the dependability of decisions or interpretations made from test scores. On the other hand, MFRM helps to identify particular elements within a facet that are problematic, or "misfitting." Through utilizing the potential of G-theory and MFRM, this study illustrated the complementary roles of these two methodologies in the validation of L2 performance assessment. In a later study, Sawaki (2007) examined the construct validity of the rating scale for a Spanish speaking assessment designed for student placement and diagnosis, using multivariate G-theory and confirmatory factor analysis (CFA) in Structural Equation Modeling (SEM). Similar to the Lynch and McNamara (1998) study, Sawaki articulated the complementary roles of the two methodologies, i.e., G-theory and CFA, in her investigation. She argued that while the G-theory could estimate and differentiate the effects of various aspects of a measurement procedure on the dependability of decisions, the CFA modeling of the rating data helped researchers examine the convergent and discriminant validity of the analytic rating scale, as well as the weighting of analytic ratings in the composite score.

These two studies clearly demonstrate how the potential of contrasting data analytic approaches might be harnessed in examining test validity. It is worth noting that such a research design also concurs with recent developments of test validity theory which advocate that multiple strands of validity evidence should be collected, evaluated, and synthesized into a validity argument to support test score

interpretation and use (e.g., Chapelle et al. 2008; Kane 2012). Equipped with the evidence generated by two different methodologies, validation researchers should be placed in a more advantageous position to interrogate the plausibility and accuracy of the warrants which are crucial to test validity, as well as the rebuttals which might weaken or undermine that validity (Kane 2012). Following this line of argument, this present study seeks to use MFRM and MTMM CFA model in SEM to examine the construct validity of the analytic rating scale of an English speaking assessment developed and used within a research university. Drawing upon the theory of interpretive validity argument (e.g., Kane 2012), this preliminary study is aimed at examining, through utilizing both MFRM and SEM, three warrants (and their respective rebuttals) which are critical to the validity of the speaking assessment: (1) Raters demonstrate sufficiently high reliability and similar severity in using the rating scale to award scores to test candidates; (2) The category structure of the rating scale functions as intended, and can effectively distinguish between test candidates at different levels of speaking proficiency; (3) Since the criteria in the rating scale represent different aspects of test candidate's L2 speaking ability (e.g., pronunciation, vocabulary, grammar), dimensions representing these aspects should be correlated, but at the same time, be distinct enough from each other. To put in another way, the test should display both convergent and discriminant validity (Sawaki 2007). Correspondingly, the three rebuttals are: (1) Raters do not demonstrate sufficiently high reliability and the same level of severity; (2) The category structure of the rating scale does not function appropriately, and thereby fails to distinguish test candidates at different levels of speaking ability; and (3) The correlations between the ability dimensions are negligible, or cannot be neatly distinguishable from each other. In this study, MFRM and SEM are used to examine these three warrants (and their respective rebuttals). Consequently, evidence in favor of these three warrants would show support for the construct validity of the rating scale, and hence the validity of the speaking assessment (Knoch 2011); conversely, lack of such evidence, i.e., evidence in favor of the rebuttals, would weaken or undermine claims for the construct validity of the rating scale.

MFRM and SEM

MFRM is a development of earlier Rasch models (e.g., dichotomous model, partial credit model) that incorporates multiple facets of the measurement procedure (Bond and Fox 2015). A facet of measurement is an aspect of the measurement procedure which the test developer claims might affect test scores, and hence needs to be investigated (Linacre 2013). Examples of such facets include the severity of rater judgments, task or item difficulty, and rating scale category options. All estimates of the measurement facets are calibrated on a single equal-interval scale (i.e., the logit scale), thereby creating a single frame of reference for interpreting the results of the analysis. Facets are estimated concurrently so they may be examined separately. Importantly, MFRM provides information about how well the performance of each

individual examinee, rater, or task matches the expected values predicted by the strict mathematical model generated during the analysis. Therefore, MFRM can help researchers detect particular elements within any facet that are "misfitting", i.e., deviating from the expectations of the mathematical model. The "misfitting" element could be a rater who is unsystematically inconsistent in applying the ratings, a task that is unexpectedly difficult, or a person whose responses are inconsistent (Lynch and McNamara 1998). In MFRM analysis, the fit statistics are calculated from the item/person residuals and are reflected in Infit and Outfit Mean Square values, both with an expected value of 1.0 (Bond and Fox 2015).

In addition to fit statistics, the MFRM analysis also reports the reliability of separation index and the separation ratio. These statistics describe the amount of variability in the measures estimated by the Rasch model for the various elements in the specified facet relative to the precision by which these measures are estimated. The reliability of separation index for each facet ranges between 0 and 1.0, whereas the separation ratio ranges from 1 to infinity (Linacre 2013). The interpretation of these two statistics, however, is different for various facets. Low separation index for the examinee facet indicates lack of variability in the examinees' ability which might be symptomatic of central tendency errors, meaning that the raters do not distinguish the performance of test candidates at different ability levels. Conversely, low values of these two statistics for the rater facet are indicative of an unusually high degree of consistency in the measures for various elements of that facet. Once parameters of the model have been estimated, interaction effects, such as the interaction between raters and rating criteria, or between raters and examinees, can be detected by examining the standardized residuals (i.e., standardized differences between the observed and expected ratings) (Eckes 2011).

Thanks to its unique advantages, MFRM has been extensively used in the fields of language assessment, educational and psychological measurement, and across the health sciences (e.g., Bond and Fox 2015; McNamara 1996; McNamara and Knoch 2012). In the field of language assessment, MFRM typically is used in rater-mediated performance assessments such as speaking or writing assessments where a score is the result of the interaction between the rater, the task, the criteria, and the examinee (Batty 2015). In particular, this analytic approach has formed the cornerstone of the descriptor scales advanced by the Common European Framework of Reference (CEFR) (e.g., North 2000; North and Jones 2009). For example, *the Manual for Relating Language Examinations to the CEFR* clearly illustrates how to use MFRM to measure the severity (or leniency) of raters, assess the degree of rater consistency, correct examinee scores for rater severity differences, examine the functioning of the rating scale, and detect the interactions between facets in writing assessment data (Eckes 2011).

In comparison with MFRM, SEM has been more widely applied for various purposes in language assessment research. Also referred to as analysis of covariance structures and causal modeling (Kunnan 1998), SEM is a comprehensive statistical methodology that "takes a confirmatory (i.e., hypothesis-testing) approach" to the analysis of a structural theory bearing on some phenomenon (Byrne 2006, p. 3), and to test theoretical hypotheses about the relationships among observed and latent

variables. It is a family of statistical techniques that includes confirmatory factor analysis, structural regression path, growth, multiple-groups, and MTMM models. The purpose of SEM is to examine whether the hypothesized relationships among variables are supported by empirical data. Usually, a model is specified *a priori* according to substantive theory, common sense, or a hypothesis to be tested. SEM is then used to estimate the discrepancy between the variance-covariance matrix as implied by the model and the observed variance-covariance matrix of the empirical data. The discrepancy is indicated by Chi-square statistics. The smaller the Chi-square value, the closer the data fit the model. In addition to the Chi-square test, a host of goodness of fit indices have been proposed to assess data/model fit, the most essential among which are CFI[1] (>0.90),[2] GFI (>0.90), SRMR (<0.05), and RMSEA (<0.05, with narrow 90 % confidence interval) (see e.g., Byrne 2006; In'nami and Koizumi 2011). When the fit is satisfactory, the model is considered to be an approximate representation of the relationships among the variables in the model. It represents one plausible explanation until future evidence falsifies this explanation (Xie and Andrews 2012).

As noted earlier, SEM techniques have been used in language assessment research for various purposes, including assessing the internal structure of a language test through structural modeling of the test data (e.g., Sawaki et al. 2009), assessing the effect of test methods on test performance (e.g., Llosa 2007), assessing equivalency of models for different populations (e.g., In'nami and Koizumi 2012), and understanding the effects of test tasks and strategy use on test performance (e.g., Kunnan 1995; Purpura 1999). SEM has also been used by language assessment researchers to investigate properties of questionnaires [see Kunnan (1998), and Ockey and Choi (2015) for a summary of the applications of SEM in language assessment research]. Despite the increasingly extensive applications of both MFRM and SEM in the field of language assessment, few attempts have been made to tap into the potential of these two different analytic approaches through combining them in test validation research. On the one hand, MFRM could function as "a magnifying glass," enabling researchers to examine closely the response patterns of individual examinees, raters, and tasks (e.g., Sawaki 2007, p. 357); SEM, on the other hand, allows researchers to hypothesize theoretical models which represent the factorial relationships between and among the variables under investigation, and to test the fit between the hypothesized model and the test data. Whereas MFRM analysis functions as the magnifying glass, SEM can provide validity evidence from a broader perspective through examining whether the hypothesized relationships between the various criteria in the rating scale are supported by the rating data, and whether such relationships are consistent with the substantive theory about language ability. Therefore, the evidence generated by

[1]CFI: Comparative Fit Index; GFI: Goodness of Fit Index; SRMR: Standardized Root Mean Residual; RMSEA: Root Mean Square Error of Approximation.

[2]The numbers in brackets are indicative of acceptable goodness of fit between the model and the empirical data.

both MFRM and SEM should be conducive to the construction of a more convincing validity argument for this speaking assessment, thus enabling us to provide a more compelling validity narrative.

Context of this Study

The Fudan English Test (FET)

In China, the College English Test (CET) has been recognized as a reliable and valid instrument in assessing university students' English language proficiency and achievement. However, recent years have witnessed the CET coming under heavy criticisms from some educators and researchers for its test format (e.g., heavy reliance on the multiple-choice questions), lack of alignment between the CET and the teaching curriculum developed within any particular university, and its rather negative washback effect on English teaching and learning at the tertiary level (e.g., Han et al. 2004). Though many of these criticisms might be seen as politically motivated or emotionally charged rather than empirically grounded, some high-ranking universities in China are attempting to develop their own English language tests in the hope of addressing the deficiencies of the CET and better aligning English testing with English teaching and learning within those university settings (see e.g., TOPE Project Team 2013; Tsinghua University Testing Team 2012). It is in this context that the FET project was initiated at Fudan University (FDU) in 2010 (see e.g., Fan and Ji 2014).

The FET is developed by the College English Center of FDU, one of the most prestigious institutions of higher learning in China. The test was formally launched in 2011, following a number of trials and pilot studies, and is currently administered once a year by FDU's Academic Affairs Office (AAO) to non-English major undergraduates. According to *the FET Test Syllabus* (FDU Testing Team 2014), the purpose of the FET is twofold: (1) to measure accurately students' English abilities and skills as reflected in the English teaching syllabus at FDU, and (2) to promote a more positive washback effect on English teaching and learning within FDU. Since September, 2011, all newly enrolled undergraduates at FDU have been required to take the FET, and to pass it within the four years of their Bachelor's program. A school-based English test notwithstanding, the FET is a reasonably high-stakes test because according to the AAO, the test is treated on a par with a compulsory English language course, which accounts for two credits in students' GPA calculations. The past few years since the inception of the FET have seen the number of test candidates increasing steadily. During the first FET administration in December 2011, 1337 students took the test,[3] and the number soared to 3575 during the most recent administration in December 2015.

[3]Typical annual undergraduate enrollment at FDU is around 3000.

Drawing upon recent models of communicative language ability and communicative language use (e.g., Bachman and Palmer 1996, 2010), the FET is designed to assess students' English language abilities in the four modalities of listening, writing, reading, and speaking, each accounting for 25 % of the test score (FDU Testing Team 2014). Previous research indicates that the FET is, on the whole, a reliable test, with internal consistency reliability coefficient reported at 0.83 (Fan and Ji 2013). Confirmatory factor analyses suggest that there is a higher-order general language competence factor and four first-order factors representing listening, reading, writing, and speaking, lending support to the construct validity of the test as well as its current score-reporting policy, i.e., reporting a composite score and four profile scores on the four subskills (Fan et al. 2014b). Furthermore, students were found to demonstrate a generally positive attitude toward the FET (Fan and Ji 2014; Fan et al. 2014a), in particular the listening, reading, and writing components. Previous research, however, has also indicated that students' attitude toward the speaking component tended to be more negative in light of the design, rating, and testing environment (Fan and Ji 2014). One concern voiced by test candidates, as previous research suggested, is that richer feedback information as to their speaking performance was lacking. An analytic scale was therefore developed to replace the holistic scale that had been used in the FET speaking component. This present study represents a preliminary attempt to validate this analytic scale developed for the FET speaking component.

The Speaking Component of the FET

The FET speaking component is a computer-mediated assessment of students' English speaking ability, administered in language laboratories. In this mode of speaking assessment, computers are used to present the tasks, and to capture students' speaking performance (Shohamy 1994). The FET speaking component comprises three tasks. In Task 1, students listen to an English passage of approximately 300 words, and respond to one or two questions based on the passage they have heard; in Task 2, students comment briefly on a topic which is mentioned in the input text; in Task 3, a graph or chart is presented on the computer screen, and students are required to describe and comment on the graph or chart. The speaking test takes about 14 min to complete. In light of the test purpose as well as previous research (e.g., Fan and Ji 2014; Fan et al. 2014a), an analytic rating scale was deemed to be more appropriate in this testing context.

The rating scale was developed on the basis of a comprehensive review of English speaking ability theory (e.g., Luoma 2004), as well as the English teaching and testing syllabus at FDU. The scale was designed to include the following four dimensions: (1) pronunciation; (2) content; (3) grammar; (4) vocabulary, all on a 4-point Likert-style scale (1-Very Poor; 2-Poor; 3-Moderate; 4-Good). Detailed descriptions accompanying each of those levels were drafted and provided. After the descriptors were drafted, they went through numerous content revisions based

on the feedback through expert reviews and panel discussions. Given the centrality of *post hoc* validity research for rating scales (e.g., Fulcher 1996; Knoch 2011), this preliminary study was conducted to examine the validity of this rating scale, and to suggest directions for its improvement in the future.

Methodology

Participants

Due to the exploratory nature of this research, convenience random sampling was employed whereby emails were sent to the prospective participants of this study, calling for their participation. Consequently, a total of 74 students participated in this study on a voluntary basis with 35 males (47.3 %) and 39 females (52.7 %). Most participants had the experience of taking the FET at least once, and therefore understood the format of the FET speaking test. To ensure that each participant was familiar with the testing procedures, a package of testing materials was sent to each of the participants one month prior to the administration, including a brief introduction to the FET speaking test, sample test papers, and marking criteria. In addition, two FET certified raters were invited to participate in this study. The two raters were both very experienced in marking the FET speaking test, and had been directly involved in the development of the analytic rating scale.

Data Collection

We used test items from the FET item bank which were written by certified item writers, and had survived earlier moderation meetings and pilot studies. Students were arranged to take the test in two language laboratories. The testing procedures simulated, as closely as possible, an authentic FET speaking test. After all test takers had completed the recordings, the two raters rated students' performance, using the analytic scale developed for this study. For the sake of data connectivity, we followed Ecke's (2011) suggestion in rating design, wherein Rater 1 rated Examinees 1–45 and Rater 2 rated Examinees 30–74 (see also Linacre 2013). Each rater was required to rate students' performance on each of the three tasks (i.e., responding to question, short comment, and graph/chart description and comment) on each of the four language aspects (i.e., pronunciation, content, grammar, and vocabulary), generating a total of 12 scores for each student (i.e., 3 tasks × 4 aspects). In total, each rater awarded 540 scores (i.e., 45 examinees × 3 tasks × 4 aspects).

Data Analysis

In this study, MFRM was first of all utilized to analyze the rating data to examine the first two warrants in relation to rater reliability and severity, and the category structure of the rating scale. Given this research scenario, a four-facet Rasch model was used which included examinee ability (74 elements), task difficulty (3 elements), rater severity (2 elements), and the difficulty of the language aspects (4 elements). The mathematical expression of this four-facet Rasch model is presented below:

$$\log\left(P_{nijmk}/P_{nijm(k-1)}\right) = B_n - D_i - C_j - T_m - F_k$$

where $P_{nijmk}/P_{nijm(k-1)}$ is the probability of examinee n receiving a rating of k in relative to $k - 1$ from rater j on criterion i for task m; B_n is the ability of examinee n; D_i is the difficulty of criterion i; C_j is the severity of rater j; T_m is the difficulty of task m; and F_k is the difficulty of receiving a rating of category k relative to immediately lower category $k - 1$. FACETS 3.71.0 (Linacre 2013) was implemented to perform MFRM analysis in this study.

To examine the third warrant about the convergent and discriminant validity of the rating scale, the MTMM CFA model in SEM was utilized to model the test data. Based on Byrne (2006) and Kunnan (1998), the SEM analytic procedures followed three steps: (1) Model specification, i.e., specifying the hypothetical MTMM models; (2) Model evaluation, i.e., evaluating the fit between the hypothesized MTMM models and the test data; and (3) Model comparison, i.e., comparing the fit of the baseline MTMM model and the alternative competing models. The SEM analysis in this study was performed with EQS 6.3 (Bentler and Wu 2005).

Results and Discussion

MFRM Analysis

Heeding the warning that "lack of connectedness among elements of a particular facet would make it impossible to calibrate all elements of that facet on the same scale" (Eckes 2011, p. 110), we first of all examined the connectedness of the resulting data set in the FACETS output. The result suggested that the rater allocation design adopted in this study was unproblematic and provided for sufficient links between all facet elements. Next, we inspected the variable map generated by the FACETS analysis. The variable map, regarded as a distinctive advantage of Rasch analysis, can illustrate graphically the estimated locations of elements in each facet on the same interval-level measurement scale, containing a wealth of basic information that is central to Rasch measurement (Bond and Fox 2015). Figure 1 displays the variable map representing the calibrations of examinees, raters, tasks,

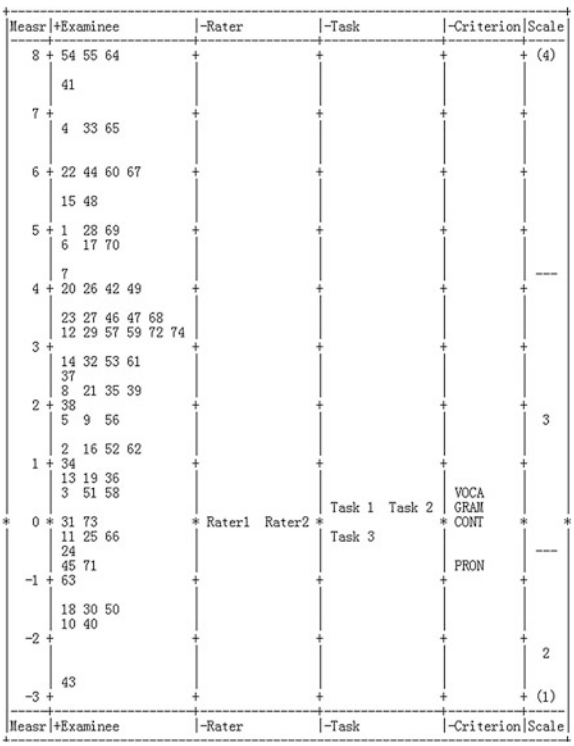


Fig. 1 The variable map

criteria, and the 4-point scale as raters used it to score examinees' performance on each language aspect. Summary statistics from the FACETS analysis for the four-facets are presented in Table 1.

Figure 1 indicated that there was a wide spread of examinees' ability with a range from −2.64 to +8.08 logits. The mean ability of examinees was 2.69 logits, with a standard error of 0.69 logits (see Table 1). The Chi-square test indicated that

Table 1 Summary statistics for the MFRM analysis

Statistics	Examinees	Raters	Tasks	Criteria
M Measure	2.69	0.01	0.26	0.60
M SE	0.69	0.00	0.00	0.00
x^2	1142.8*	0.01	8.9*	50.1*
df	73	1	2	3
Separation index	3.64	0.00	1.88	4.03
Separation reliability	0.93	0.00	0.78	0.94

Note *Significant at the $p > 0.05$ level

the examinees came from statistically distinct ability groups (x^2 = 1142.8, df = 73, $p < 0.01$). Figure 1 also revealed that most examinees were located above the difficulty of the three speaking tasks in the variable map. The mean ability of examinees (2.69 logits) was substantially higher than the mean task difficulty (0.26 logits), suggesting that, on average, the three tasks were quite easy for this group of test candidates. It should be noted, however, that all participants in this study were volunteers, and students at higher ability levels might be more motivated to participate in such a study. That said, the results indicate that more difficult tasks might be developed in the future to tap into test candidates' speaking ability. The satisfactorily high reliability (0.93) indicated the reproducibility of the measures, suggesting that the same number of statistically distinct levels of proficiency could be expected if we repeated the same data collection (Linacre 2013).

Of particular interest to this current study is the rater facet. The interpretation of the statistics for raters in Table 1, however, is decidedly different from that for the other three-facets. When raters within a group exercised a highly similar degree of severity, rater separation reliability will be close to 0 (Eckes 2011). This is exactly what happened in this study where the two raters were found to demonstrate highly similar patterns in their rating behavior. This could be first observed from Column 3 of the variable map, as shown in Fig. 1. In addition, this was also indicated by the insignificant Chi-square test, as well as the extremely low separation index and reliability (see Table 1). Rater fit statistics present statistical indicators of the degree to which raters used the rating scale in a consistent manner (Eckes 2011). The Infit and Outfit Mean Square were 1.03 and 1.09 for Rater 1, and 0.94 and 0.92 for Rater 2, all approximating the ideal value of 1. These statistics suggested that both raters were consistent in their ratings. Such a finding is unlikely when multiple raters are used, and is at odds with previous research which tended to identify significant rater effects (e.g., Eckes 2005; Lynch and McNamara 1998). As a preliminary study, only two raters were involved; both raters, as noted earlier, were very experienced in rating the FET speaking test, and were directly involved in the construction of the rating scale. As such, caution needed to be exercised in overinterpreting the results emanating from this part of the research. A larger and more representative sample of raters should be included in a future investigation. On the basis of this preliminary study, it seems reasonable to conclude that the first warrant, i.e., the two raters demonstrate sufficiently high reliability and similar level of severity, was supported.

The second warrant is pertinent to the utility of the category structure of the rating scale. To verify the functioning of each response category, Linacre's (2004) criteria were applied, including: (1) A minimum of 10 observations is needed for each category; (2) Average category measures must increase monotonically with categories; (3) Outfit Mean Square statistics should be less than 2.00; (4) The category threshold should increase monotonically with categories; (5) Category thresholds should be at least 1.4–5-logits apart, and (6) The shape of the probability curves should peak for each category (cited in Oon and Subramaniam 2011, p. 125). Summary of category structure of the 4-point scale is presented in Table 2. As shown in this table, though all categories were used by raters, the first category (i.e., Very Poor) was substantially under-used with only 1 % frequency, suggesting

Table 2 Summary of category structure of the 4-point rating scale

Category	Observed count (%)	Average measure	Outfit MnSq	Threshold calibration
1. Very Poor	14 (1 %)	−1.99	0.80	None
2. Poor	137 (14 %)	−0.65	0.90	−3.55
3. Moderate	538 (57 %)	2.10	1.00	−0.73
4. Good	259 (27 %)	4.96	1.10	4.18

that this category should be removed or collapsed with its adjacent category (Bond and Fox 2015). It should be noted that this finding concurred with our earlier observation that the three tasks in this speaking test were, on average, too easy for this sample of test candidates. The "Very Poor" category should be attempted on a larger and more representative sample of test candidates in the future to further examine the functioning of this category. Table 2 also showed that average category measures increased monotonically from −1.99 to 4.96, suggesting that these categories were used as expected by the raters. Outfit Mean Square values ranged from 0.80 to 1.10, suggesting that these categories did not introduce noise into the measurement process. An inspection of the distance between two adjacent categories showed that the required range of 1.4–5-logits was met, suggesting that the four categories defined distinct positions on the latent variable. The category probability curves (displayed in Fig. 2) further revealed that each category emerged as a peak. The analysis of the category structure only partly supported the second warrant, and suggests that the category structure should be revised in the future through either removing the redundant category or collapsing it with its adjacent category.

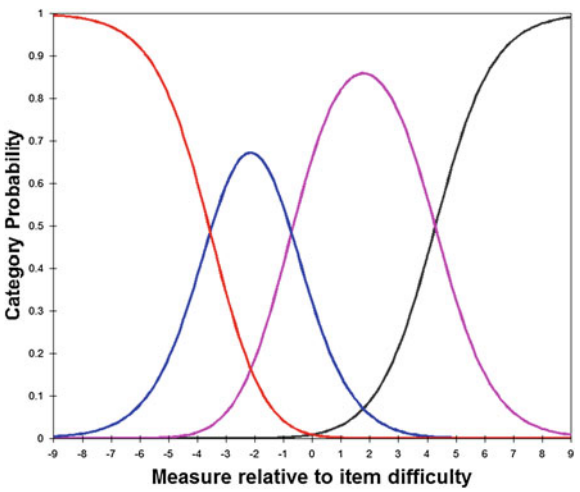


Fig. 2 Category probability curves for the 4-point rating scale

SEM Analysis

The SEM analysis in this study followed the procedures outlined by Byrne (2006). In this analysis, a MTMM design was adopted by which multiple traits are measured by multiple methods. The four language performance aspects (i.e., pronunciation, content, grammar, and vocabulary) were specified as the trait factors in the MTMM model, whereas the three tasks (i.e., responding to questions, short comment, and graph/chart description and comment) were the method factors. According to Campbell and Fiske (1959), convergent validity refers to the extent to which different assessment methods concur in their measurement of the same trait, whereas discriminant validity refers to the extent to which independent assessment methods diverge in their measurement of different traits. The convergent and discriminant validity of the rating scale could be examined at both the matrix and parameter level, as advised by Byrne (2006). Specifically, four MTMM CFA models were specified in this study, including Correlated Traits/Correlated Methods Model (Model 1), No Traits/Correlated Methods Model (Model 2), Perfectly Correlated Traits/Freely Correlated Methods Model (Model 3), and Freely Correlated Traits/Uncorrelated Methods Model (Model 4). Readers are referred to Figs. 3, 4, 5, and 6 for the graphic representations of these four MTMM models. It is worthnoting that: (1) the variances of latent factors in the four models were set to be 1.0 for model identification purposes; (2) the only difference between Model 1 and Model 3 (displayed in Figs. 3 and 5 respectively) lies in that the correlations between the trait factors in Model 1 were freely estimated, but fixed to 1 in Model 3, as indicated by the dotted lines in the figure.

Among the four hypothesized models, Model 1 was the least restrictive model, and therefore served as the baseline model against which the alternative MTMM

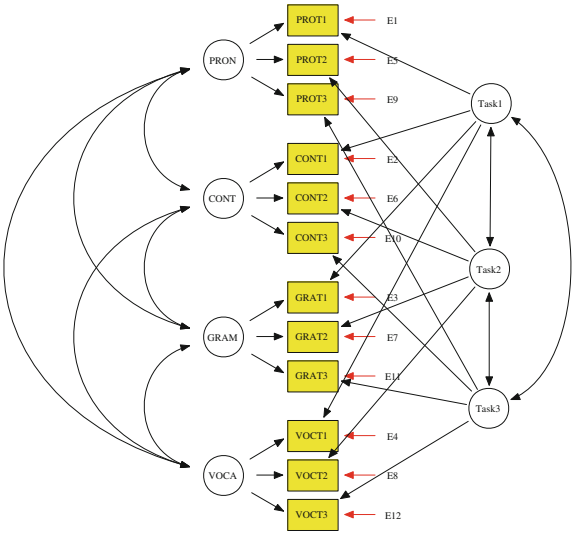


Fig. 3 Correlated Traits/Correlated Methods Model (Model 1)

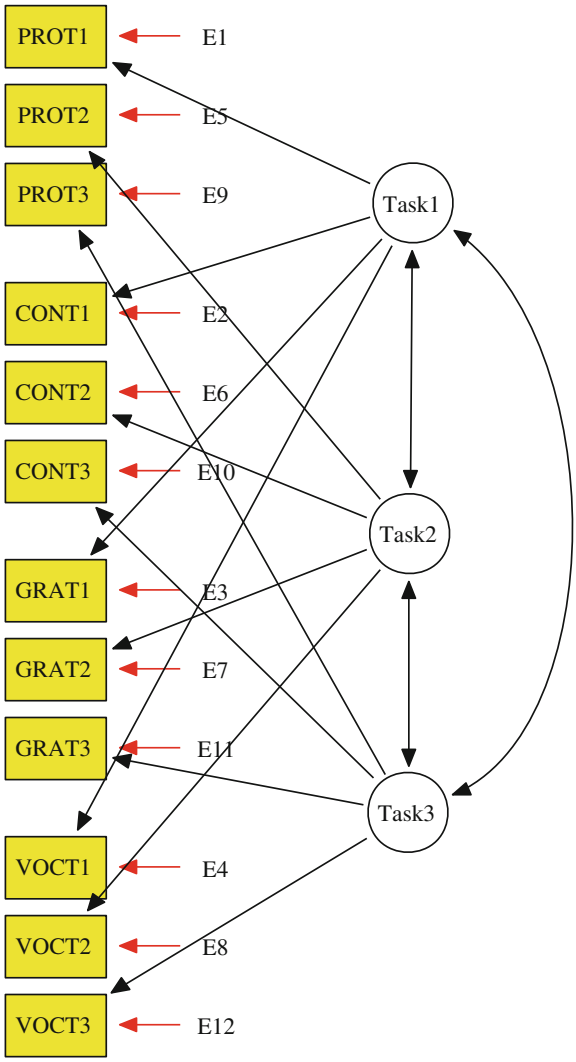


Fig. 4 No Traits/Correlated Methods Model (Model 2)

models were compared. Since the other three models were nested models of Model 1, the Chi-square difference test was used to compare whether the difference between the fit of the alternative models and the baseline model was statistically significant. A non-significant Chi-square value would indicate that the difference was negligible. In addition to the Chi-square difference test, Cheung and Rensvold (2002) recommended that if the CFI difference values did not exceed 0.01, then the difference between the fit of the two models would be of minimal practical significance. While aware that the CFA format allows for an assessment of construct validity at both matrix and individual parameter levels (Byrne 2006), this

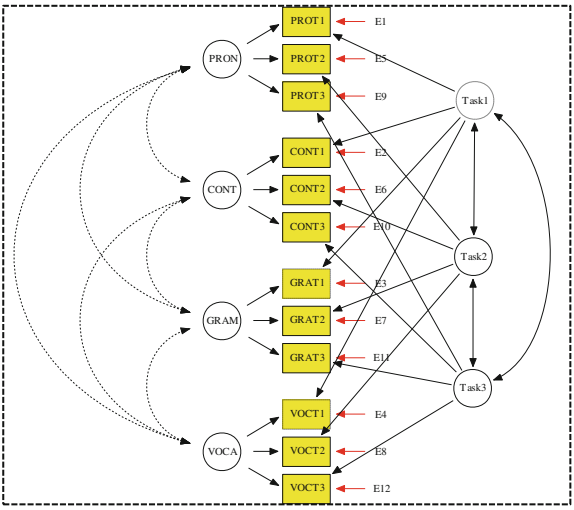


Fig. 5 Perfectly Correlated Traits/Correlated Methods Model (Model 3)

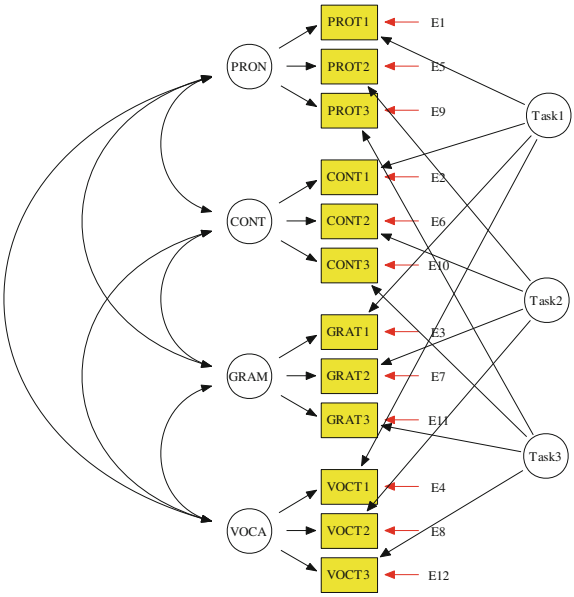


Fig. 6 Freely Correlated Traits/Uncorrelated Methods Model (Model 4)

preliminary study tested for evidence in relation to convergent and discriminant validity primarily at the matrix level.

Given that Mardia's normalized estimate was 3.74, the data were considered normally distributed, and hence the default estimation method in EQS, i.e., the maximum likelihood method, was used for parameter estimation purposes (Bentler and Wu 2005). The goodness-of-fit indexes of the four MTMM models are

Table 3 Summary of Goodness-of-Fit Indexes for MTMM Models

Model	x^2	*df*	CFI	GFI	SRMR	RMSEA (90 % C.I.)
Model 1	35.93	33	0.995	0.934	0.038	0.035 (0.000–0.094)
Model 2	142.59*	51	0.858	0.756	0.083	0.187 (0.126–0.186)
Model 3	59.91*	39	0.968	0.890	0.039	0.086 (0.037–0.126)
Model 4	42.44	36	0.990	0.912	0.047	0.050 (0.000–0.101)

Notes Model 1: Correlated Traits/Correlated Methods Model; Model 2: No Traits/Correlated Methods Model; Model 3: Perfectly Correlated Traits/Correlated Methods Model; Model 4: Freely Correlated Traits/Uncorrelated Methods Model. *$p < 0.05$

presented in Table 3. As could be seen from this table, the baseline model, i.e., the Correlated Trait/Correlated Method model fits the data well [$x^2 = 35.93$, $df = 33$, $p > 0.05$, CFI = 0.995, RSMEA = 0.035 (90 % C.I., 0.000–0.094)]. In comparison, Model 2, i.e., No Traits/Correlated Methods Model, displayed extremely poor fit to the data [$x^2 = 142.59$, $df = 51$, $p < 0.05$, CFI = 0.858, RSMEA = 0.187 (90 % C.I., 0.126–0.186)]. A Chi-square difference test was performed to compare the fit of these two models, yielding a highly significant result ($\Delta x^2 = 106.66$, $df = 18$, $p < 0.01$). Furthermore, the CFI difference value was 0.137, well above the criterion value of 0.01 recommended by Cheung and Rensvold (2002). The results supported the convergent validity of the rating scale which required correlations between independent measures of the same trait (e.g., the ratings of pronunciation on Task 1, Task 2 and Task 3) that should be substantial and statistically significant.

Discriminant validity, on the other hand, is assessed in terms of both traits and methods. In testing for evidence of discriminant validity among traits, a model in which trait factors were posited to be freely estimated (Model 1) was compared with one in which they were perfectly correlated (Model 3). An inspection of the goodness-of-fit indexes in Table 3 revealed that Model 3 did not fit the data well [$x^2 = 59.91$, $df = 39$, $p < 0.05$, CFI = 0.968, RSMEA = 0.086 (90 % C.I., 0.037–0.126)]. The comparison between Model 1 and Model 3 yielded a statistically significant result ($\Delta x^2 = 23.98$, $df = 6$, $p < 0.05$), and the difference in practical fit was quite large (ΔCFI = 0.03). This result supported discriminant validity among traits, and suggested that although the traits were substantially correlated, they were still distinguishable from each other. In the field of language assessment, numerous factor analytic studies have supported the notion that language ability is a complex construct with multiple dimensions, though the research community has not reached an agreement regarding the nature of the constituents, or on the manner in which they interact (e.g., Gu 2014; In'nami and Koizumi 2012; Sawaki et al. 2009). The tenability of Model 1 and rejection of Model 3 lends further support to this view, suggesting that not only is general language ability multidimensional, but any single language modality such as speaking ability might also have multiple constituents.

Based on the same logic, when testing for the evidence of discriminant validity related to method effects, a model in which method factors were posited to be freely estimated (Model 1) was compared with one in which method factors were

specified to be uncorrelated (Model 4). The goodness-of-fit indexes in Table 3 indicated that Model 4 was a reasonably satisfactory fit to the data [$\Delta x^2 = 42.44$, $df = 36$, $p > 0.05$, CFI = 0.912, RSMEA = 0.050 (90 % C.I., 0.000–0.101)]. A comparison of this model with the baseline model (i.e. Model 1) yielded a Δx^2 value which was statistically not significant ($x^2 = 6.51$, $df = 3$, $p > 0.05$) with negligible difference in CFI values (ΔCFI = 0.01). A large Δx^2 or substantial ΔCFI argued for the lack of discriminant validity, thereby suggesting common method bias. Given that this analysis yielded a non-significant Chi-square test and the ΔCFI was minimal, it was reasonable to conclude that the scale displayed evidence of discriminant validity related to methods.

An inspection of the factor loadings in the baseline model, i.e., the Correlated Traits/Correlated Methods Model revealed that the path coefficients of the observed ratings to the corresponding four trait factors in the model were high and significantly different from zero, ranging from 0.48 to 0.89. The results indicated strong linear relationships between the trait factors and the observed ratings. In addition, the correlations between the four trait factors were significant, ranging from 0.49 to 0.93. The substantial path and correlation coefficients again support the convergent validity of the rating scale related to the traits. However, the correlations between the three method factors were found to be reasonably high. For example, the correlation between Task 2 and Task 3 was 0.56. The high correlation coefficients between the methods argued against discriminant validity in relation to the methods, and suggested common method bias in measurement (Byrne 2006). These results recommended that the FET speaking test designers should adopt elicitation methods which are distinct enough so as to avoid common method bias. Taken together, the MTMM modeling of the test data lent reasonably strong support to the third warrant, i.e., the rating scale displays both convergent and discriminant validity. However, the strength of this warrant was somewhat weakened by the identification of common method bias which should be addressed in future revisions of this speaking assessment.

Conclusions, Limitations, and Implications

This preliminary validation research demonstrated how MFRM and SEM could be used in tandem in the interrogation of the construct validity of the analytic rating scale developed for a school-based English speaking assessment. Through harnessing the potential of both research methodologies, this study examined the plausibility and accuracy of three warrants (and their respective rebuttals) which were deemed crucial to the construct validity of the rating scale, and hence to this speaking assessment. Specifically, a four-facet MFRM model was utilized to calibrate examinee ability, rater severity, task difficulty, and the difficulty of ability dimensions on the same interval measurement scale. MFRM analysis of the rating data lent support to the first warrant, i.e., raters displayed high reliability and the same level of severity in awarding scores to test candidates. The second warrant,

i.e., the category structure of the rating scale functioned appropriately, was partly supported by the MFRM analysis. The lowest category was found to be substantially under-used, thereby weakening the strength of this warrant. The third warrant regarding the convergent and discriminant validity of the rating scale was examined through using the MTMM CFA design to model the rating data. Four MTMM models were specified, evaluated, and compared. As it turned out, the SEM analysis partly supported the third warrant, suggesting that the rating scale displayed convergent and discriminant validity in relation to both traits and methods. The high correlations between the method factors, however, argued against the discriminant validity about methods, and suggested common method bias. This finding somewhat weakened the strength of the third warrant, and should be addressed in future test revisions.

In this research, the Rasch model and SEM have been used on the same set of data, but separately. Bond and Fox (2015) are much more direct about using the models collaboratively, in conjunction: "For those who are more thoughtfully wedded to SEM, our advice would be spread over two steps: First, that Rasch analysis should be adopted to guide the construction and quality control of measurement scales for each of the variables that feature in the research. Second, the interval-level person Rasch measures and their standard errors (SEs) that derive from each of those instruments should be imputed into the SEM software for the calculation of the relationships between those variable measures" (p. 240). In defense of the current analyses and results, we assert that this speaking test has almost satisfied Rasch model requirements for the production of interval-level measurement. To the extent that the data fit the Rasch model, total scores are the necessary statistic for parameter estimation, so using raw ordinal-level data in the SEM analyses is likely to be unproblematic in this case. For researchers who wish to explore the applicability of further developments of the Rasch model for answering the questions broached in this research, a generalized form of the Rasch model, the mixed coefficients multinomial logit model (MCMLM; Adams et al. 1997) might be applied. Such multidimensional item response model analyses combine the response information for different tests according to the size of the correlations between the latent variables. When the correlations are high but not perfect, as in this case, the MIRM uses information from all tests to estimate performances on each of the latent traits (after Bond and Fox 2015, pp. 291–292).

The research described herein has several limitations. First, as a preliminary study, convenience sampling was adopted. Consequently, it cannot be held that the sample of test candidates used in this study was representative of the test population for which this test was designed. Moreover, SEM is a large-sample analytic technique (e.g., In'nami and Koizumi 2011; Kline 2005). Given the complexity of the MTMM models specified in this study, a larger sample of test candidates is essential for ensuring the viability of parameter estimations. Also, in view of the central role that raters played in performance assessment, a larger and more representative rater sample should be attempted in future validation research. Second, some essential features of the MFRM analysis, such as the interaction or bias analysis (e.g., Eckes 2011; Linacre 2013) were not included in this research. Investigations into the

potential interactions between raters and the criteria in the rating scale could be particularly meaningful to such a study. Finally, the MTMM CFA format allows the researchers to investigate the convergent and discriminant validity at both matrix and parameter levels (Byrne 2006). This preliminary study, however, was primarily focused on evidence at the matrix level. A closer examination of convergent and discriminant validity at the parameter level could be very revealing, and should therefore be attempted in the future. Meanwhile, SEM allows researchers to hypothesize and evaluate a host of different theoretical models. Some alternative MTMM models could therefore be subsequently specified and evaluated, such as the Higher-Order Trait Model (e.g., Sawaki 2007) with a view to understanding more clearly the relationships between and among the observed and latent variables in this study.

This research has implications for the future revision and improvement of the analytic rating scale, as well as the speaking assessment under study. First, the category structure of this rating scale warrants adjustment. The redundant category, as discussed earlier, could be either removed or collapsed with its adjacent category. Second, the tasks in the speaking assessment could be redesigned. Given the common method bias identified by SEM analysis, testing methods which are sufficiently distinguishable from each other could be adopted in the future. In the current test format, both Task 2 and Task 3 in the FET speaking component require test candidates to give comments on a certain topic; such a design is very likely to cause common method bias. Other task formats such as reading aloud or listening to summarize could be attempted in the future development of this speaking test (see e.g., Fan 2014). Finally, this study has methodological implications for language assessment researchers, in particular the developers and validators of performance assessments (e.g., speaking, writing). Echoing the view of Bond and Fox (2015) regarding the collaborative use of Rasch model and SEM in conjunction, future researchers may consider using the MFRM to examine the quality of the rating scale, and revise it accordingly before utilizing the SEM to either test the tenability of specific theoretical models or examine the convergent and discriminant validity at both matrix and parameter levels. By doing so, the potential of the two methodologies could be harnessed more adequately.

Acknowledgments The study reported in this chapter was supported by the National Social Sciences Fund of the People's Republic of China under the project title of "Development and Validation of Standards in Language Testing" (Grant No: 13CYY032), and the Research Project of National Foreign Language Teaching in Higher Education under the project title of "Teacher-, Peer-, and Self-assessment in Translation Teaching: A Many-Facets Rasch Modeling Approach" (Grant No: 2014SH0008A). Part of this research was published in the third issue of *Foreign Language Education in China* (*Quarterly*) in 2015.

References

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–24.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238–257.

Bachman, L. F., & Palmer, A. S. (1996). *Language assessment in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Batty, A. O. (2015). A comparison of video-and audio-mediated listening tests with Many-Facets Rasch modeling and differential distractor functioning. *Language Testing, 32*(1), 3–20.

Bentler, P. M., & Wu, E. J. (2005). *EQS 6.1 for Windows*. Encino, CA: Multivariate Software.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*: New York: Routledge.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, New Jersey: Psychology Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York and London: Routledge, Taylor & Francis Group.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facets Rasch analysis. *Language Assessment Quarterly: An International Journal, 2*(3), 197–221.

Eckes, T. (2011). *Introduction to many-facets Rasch measurement*. Frankfurt: Peter Lang.

Fan, J. (2014). Chinese test takers' attitudes towards the Versant English Test: A mixed-methods approach. *Language Testing in Asia, 4*(6), 1–17.

Fan, J., & Ji, P. (2013). Exploring the validity of the Fudan English Test (FET): Test data analysis. *Foreign Language Testing and Teaching, 3*(2), 45–53.

Fan, J., & Ji, P. (2014). Test candidates' attitudes and their test performance: The case of the Fudan English Test. *University of Sydney Papers in TESOL, 9*, 1–35.

Fan, J., Ji, P., & Song, X. (2014a). Washback of university-based English language tests on students' learning: A case study. *The Asian Journal of Applied Linguistics, 1*(2), 178–192.

Fan, J., Ji, P., & Yu, L. (2014b). Another perspective on language test validation: The factor structure of language tests. *Theory and Practice in Foreign Language Teaching, 4*, 34–40.

FDU Testing Team. (2014). *The FET Test Syllabus*. Shanghai: Fudan University Press.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing, 31*(1), 111–133.

Han, B., Dan, M., & Yang, L. (2004). Problems with College English Test as emerged from a survey. *Foreign Languages and Their Teaching, 179*(2), 17–23.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEFL test: A multi-sample analysis. *Language Testing, 29*(1), 131–152.

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly, 8*(3), 250–276.

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81–96.
Kondo-Brown, K. (2002). A FACET analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*, 3–31.
Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling approach* Cambridge: Cambridge University Press.
Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing, 15*(3), 295–332.
Linacre, M. (2013). *A user's guide to FACETS (3.71.0)*. Chicago: MESA Press.
Linacre, M. (2004). *Optimal rating scale category effectiveness*. In E. V. Smith & R. M. Smith (Eds.), Introduction to Rasch measurement (pp. 258–278). Maple Grove, MN: JAM Press.
Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*(4), 489–515.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.
Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facets Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158–180.
McNamara, T. (1996). *Measuring second language proficiency*. London: Longman.
McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. Language Testing, *29*(4), 553–574.
North, B. (2000). *The development of common framework scale of language proficiency*. New York: Peter Lang.
North, B., & Jones, N. (2009). *Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling*. Strasbourg: Language Policy Division.
Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, *12*(3), 305–319.
Oon, P. T., & Subramaniam, R. (2011). Rasch modelling of a scale that explores the take-up of Physics among school students from the perspective of teachers. In R. F. Cavanaugh & R. F. Waugh (Eds.), *Applications of Rasch measurement in learning environments research* (pp. 119–139). Netherlands: Sense Publishers.
Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.
Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing, 16*(4), 457–478.
Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing, 29*(2), 223–241.
Sawaki, Y. (2007). Construct validation of analytic rating scale in speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355–390.
Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 5–30.
Shin, S.-Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing, 32*(2), 259–281.
Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2), 99–123.
TOPE Project Team. (2013). *Syllabus for Test of Oral Proficiency in English (TOPE)*. Beijing: China Renming University Press.
Tsinghua University Testing Team. (2012). *Syllabus for Tsinghua English Proficiency Test (TEPT)*. Beijing: Tsinghua University Press.
Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–12.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing, 16*(1), 82–111.

Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*(1), 49–70.

A Rasch Model Analysis of the "Four L2 Anxieties"

Matthew T. Apple

Introduction

Foreign or second language (L2) anxiety is considered as a situation-specific variable that arises within foreign language contexts, related to but separate from general trait/state anxiety and test anxiety. However, existing instruments that measure L2 anxiety have not been examined for item fit or unidimensionality and frequently produce different results from study to study, even within similar contexts. In addition, claims that different language-skill related L2 anxieties are independent have been based on traditional forms of statistical investigation that rely on correlational analysis to determine construct validity. The lack of valid unidimensional measurement instruments poses a problem for L2 teachers who need to identify sources of anxiety for their students. The present study aims to investigate (1) the construct validity of instruments measuring the "four L2 anxieties" and test anxiety, and (2) to what degree the four anxieties and testing anxiety are distinct constructs.

Background of the Issue

Definitions of Anxiety

The concept of anxiety as both a personality trait as well as a temporary state was established almost half a century ago and has been measured traditionally by the State-Trait Anxiety Inventory (STAI, Spielberger 1983). Gardner (1985) was one of

M.T. Apple (✉)
Ritsumeikan University, Kyoto, Japan
e-mail: mapple@fc.ritsumei.ac.jp

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_4

the first second language acquisition (SLA) researchers to develop an instrument solely designed to measure foreign language anxiety—a short five-item scale called "French classroom anxiety"—but it was the study by Horwitz et al. (1986) that popularized the term "foreign language anxiety" for L2 teaching. Horwitz et al. argued that L2 anxiety in the classroom was composed of three primary components—communication apprehension, fear of negative evaluation by the teacher, and test anxiety—and created the Foreign Language Classroom Anxiety Scale (FLCA), with 33 positively and negatively worded items.

Based on a series of studies conducted in the late 1980s and early 1990s (e.g., MacIntyre and Gardner 1989, 1991) L2 anxiety has been viewed as a "situation-specific" anxiety, e.g., a type of state anxiety that occurs only when using one's L2 in certain social situations. Based on factor analysis, MacIntyre and Gardner (1991) argued that test anxiety and "general anxiety" (which they defined based on a combination of items concerning L1 usage in various social situations) differed from L2 anxiety. Other studies since then have proposed L2 skills-related anxieties and created new L2 anxiety instruments, such L2 speaking anxiety (Apple 2013), L2 listening anxiety (Kim 2000), L2 reading anxiety (Saito et al. 1999), and L2 writing anxiety (Cheng et al. 1999). However, problems remain in SLA anxiety research. Dewaele (2013) pointed out that existing L2 anxiety instruments do not correlate well, and the assumption that L2 anxiety is independent of trait anxiety is somewhat suspect. In fact, evidence that L2 anxiety is strongly related to first language (L1) anxiety and trait anxiety has already been attested in communication studies (Jung and McCroskey 2004), though seemingly ignored by SLA researchers. In addition, the language used in L2 anxiety measurement instruments, compounded by a lack of item fit analysis and construct validity, is a problem for cross-sample validity of L2 anxiety instruments.

Conceptual Problems

A major issue concerning L2 anxiety is the very language used to describe it. As previously mentioned, the "trait" versus "state" distinction has been suggested. However, previous research has cast doubt upon the validity of the STAI (e.g., Kaipper et al. 2010; Tenenbaum et al. 1985). The terms "facilitative" as well as "debilitative" have been used to describe types of L2 anxiety, based on the idea that anxiety can both help and hinder language performance (e.g., Price 1991; Young 1999). However, the items used to measure L2 anxiety contain words that may or may not pertain to the measurement of anxiety, for example: afraid, annoyed, concerned, frightened, frustrated, nervous, panic, tense, uncertain, uncomfortable, uneasy, upset, and worried. Compared to a dictionary-definition of anxiety (Fig. 1),

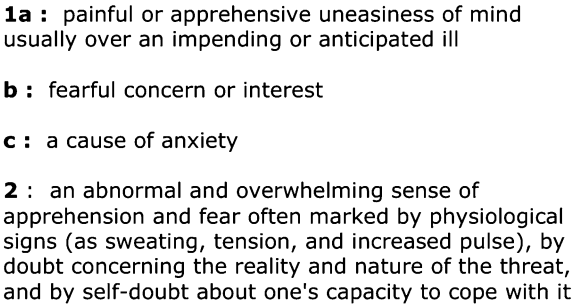


Fig. 1 Anxiety definitions from Merriam-Webster (*Source* http://m-w.com)

these words sometimes exemplify and sometimes are only tangetially related with traditional ideas of what constitutes "anxiety." Key components missing from L2 anxiety measurement instruments are the anticipation of something painful and the idea that anxiety is abnormal. Physical descriptions (e.g., sweating, increased heart rate) are also typically missing. Thus, even before considering measurement problems, there are already conceptual, and thus content validity, concerns in L2 anxiety instruments.

Measurement Problems

In addition to content validity, an overriding concern regarding the use of Likert-scale questionnaire instruments to measure L2 anxiety is construct validity, without which there is no way of knowing whether the results obtained tell us that the construct has actually been measured rather than the result of errors, disturbances, and other random statistical noise. An additional requirement of the L2 anxiety measurement instruments that has not yet been confirmed is unidimensionality; each L2 anxiety measurement instrument should measure only one construct, rather than several related constructs. L2 anxiety researchers who use multidimensional construct instruments cannot be certain which data come from which constructs, making generalization of their findings extremely difficult, if not impossible.

Traditionally, L2 anxiety measurement instruments have been "validated" through correlation to existing instruments and split-half reliability analysis using Cronbach's alpha. However, neither correlational analysis nor Cronbach's alpha measure construct validity, facts which have been known for nearly half a century (Green et al. 1977; Cortina 1993). Excepting for L2 speaking anxiety (Apple 2013), the separate L2 instruments have not been subject to item fit analysis and unidimensionality of construct has not been validated. Finally, the relationship among the four L2-skills anxieties and testing anxiety still need to be clarified.

Methods

Participants

This study comprised 315 first and second-year students in 12 intact EFL classes at two undergraduate universities in Kyoto, Japan. Fourteen questionnaires were discarded due to incomplete answers, for a total sample size of $N = 298$. There were 150 male and 148 female students; 126 were first year students and 172 were second-year students. Study participants had an intermediate English proficiency level as measured by the TOEIC ($M = 569.34$, SD = 108.67) and comprised three majors: social sciences (30.9 %), letters/humanities (39.3 %) and economics (29.9 %).

Instruments

A Likert-type questionnaire instrument[1] with six points (1 = strongly disagree, 6 = strongly agree) was used, with 51 items chosen from five existing L2 anxiety instruments:

1. L2 speaking anxiety ($k = 11$; Apple 2013).
2. L2 listening anxiety ($k = 11$; Kimura 2008).
3. L2 reading anxiety ($k = 11$; Matsuda and Gobel 2004).
4. L2 writing anxiety ($k = 10$; Cheng et al. 1999).
5. Test anxiety ($k = 8$; In'nami 2006).

Of the five instruments, only one (Apple 2013) had been created and analyzed using Rasch Analysis prior to the study. Aside from Apple (2013), the original instruments were constructed according to traditional statistic-based methods, i.e., using large numbers of items to drive up the Cronbach's alpha estimate and analyzed according to factor analysis or correlational analysis using Pearson's coefficient. L2 Listening Anxiety originally consisted of 33 items. L2 Reading Anxiety originally consisted of 20 items.[2] L2 Writing Anxiety originally comprised 25 items. Testing Anxiety was originally a combined questionnaire with items from two separate instruments totalling 57 items (see Discussion for more on this instrument). In the interests of reducing questionnaire fatigue, only the top 11 items (<0.40) from each instrument were selected based on existing traditional factor analysis (EFA) factor loadings.

Procedures

A bilingual English–Japanese questionnaire was distributed via SurveyMonkey, an online survey tool that allowed study participants to complete the survey using

smartphone-based browser software. Three L1 teachers of English assisted the researcher in collecting data; teachers were asked to read the directions out loud (in English) to students, encouraged students to select responses honestly and without discussing with their classmates, and informed students that the survey was voluntary and would not affect their course grades. No student chose to opt out of the survey. Data obtained from the questionnaire were analyzed in WinSteps 3.75 using the Rasch rating scale model (Andrich 1978), which is a polytomous data form of the Rasch model (Rasch 1960). To judge Lickert category functioning, the four criterion of Linacre (2002) were used: (1) At least 10 observations should be present for each step of the scale, (2) average person measures for each step should be higher than the average person measures of the previous step, (3) outfit mean squares of each step should be less than 2.0, and (4) gaps in step difficulties should be no fewer than 0.59 and no greater than 5 logits.

The criterion to determine item fit were mean squares from 0.7 to 1.3 (Bond and Fox 2007), with 1.0 denoting a perfect fit to the Rasch model's expectations. Item-person maps were additionally requested as visual confirmation of the constructs. Rasch principal component analysis (PCA) of construct residuals were also conducted to check unidimensionality of the individual constructs using the technique known as Rasch factor analysis (Wright 1996). The criteria used to determine construct validity were 50 % of variance accounted for by the primary dimension (the Rasch model), with secondary dimension contrasts accounting for 10 % or less variance and having less than 3.0 eigenvalues (Linacre 2007). The combination of item fit analysis and Rasch PCA can thus be used to support claims of cross-sample generalization (Wolfe and Smith 2007).

Results

As a preliminary step, all items were simultaneously input into the Rasch model and an item-person map was requested to provide a visual image of the overall targetting of the items (Fig. 2). In general, the items were appropriately targeted for the sample population; however, the item endorsability range tended to clump around the mean of the population and there were several participants whose anxiety levels could not be adequately measured by the items. Writing anxiety items (wa) tended to be more difficult to endorse, and listening anxiety items (la) tended to be easier to endorse compared to the other three anxieties (speaking, reading, testing), whose items were more evenly distributed. Rasch item reliability was 0.98 and item separation was 6.64, and Rasch person reliability was 0.96 and person separation was 4.76 for all items.

Since the various anxiety instruments were originally meant to measure a single construct, items from each instrument were input into Rasch model. Likert category analysis was conducted to determine the appropriate usage of the six Likert categories prior to item fit analysis and Rasch PCA for individual constructs.

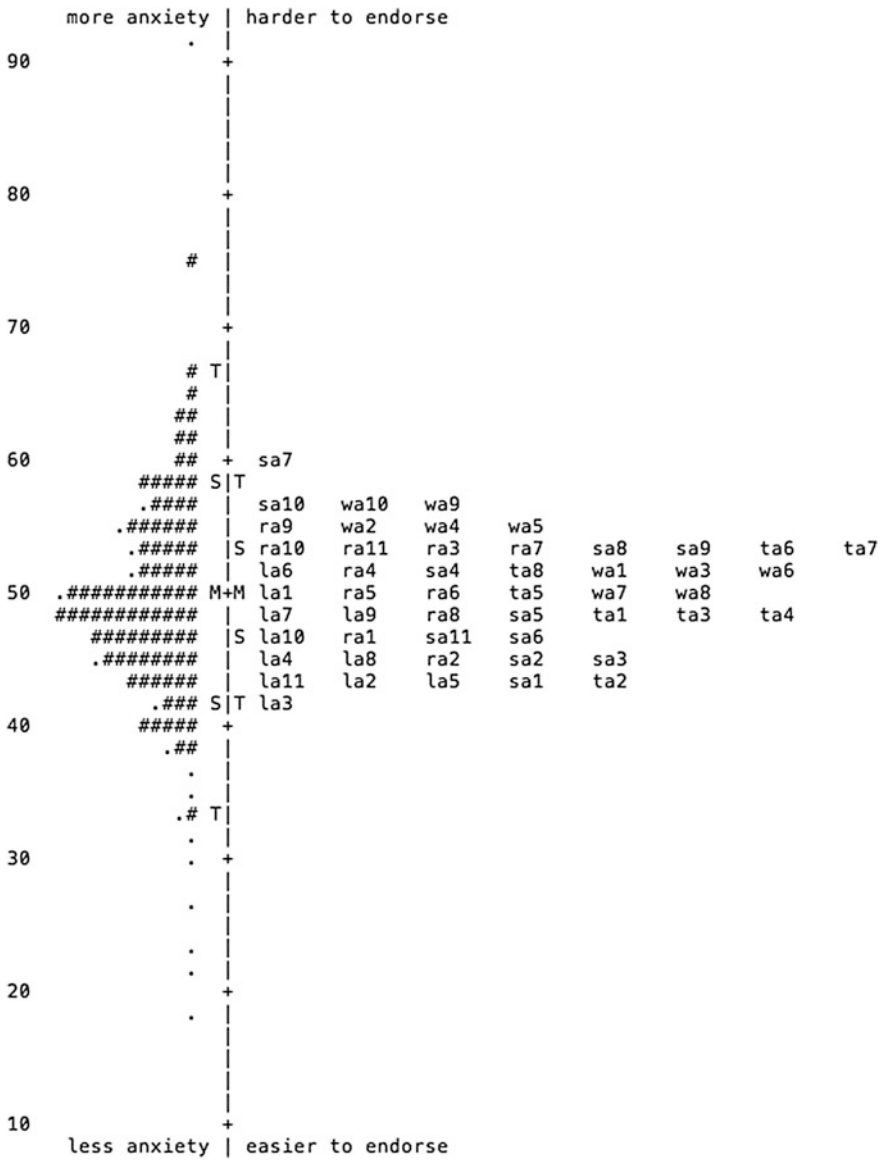


Fig. 2 All 51 items from the five anxiety-related instruments, prior to item fit analysis. *M* stands for mean. *S* stands for one standard deviation from the mean. *T* stands for two standard deviations from the mean; sa = speaking anxiety; la = listening anxiety; wa = writing anxiety; ra = reading anxiety; ta = testing anxiety; $N = 298$

Likert Category Utility and Item Misfit Analysis

Likert category utility analysis showed misfit for the first category in L2 listening anxiety and for the sixth category in L2 writing anxiety, L2 reading anxiety, and Testing anxiety.[3] No Likert categories misfit in L2 speaking anxiety. Likert categories 1 and 2 were combined for L2 listening anxiety, and Likert categories 5 and 6 were combined for L2 writing anxiety, L2 reading anxiety, and Testing anxiety prior to item fit analysis.

Rasch model item analysis revealed that seven items misfit their intended constructs (Table 1). Item Sa1 ("I'm worried that other students in class speak English better than I do") misfit the L2 speaking anxiety construct (infit MNSQ = 1.81, outfit MNSQ = 1.90). Item La1 ("I get stuck with one or two unfamiliar words") misfit the L2 Listening Anxiety construct (infit MNSQ = 1.32, outfit MNSQ = 1.40). Two items misfit the L2 reading anxiety construct, item Ra8 ("It bothers me to encounter words I can't pronounce while reading English," infit MNSQ = 1.88, outfit MNSQ = 1.85) and item Ra10 ("By the time you get past the funny letters and symbols in English, it is hard to remember what you are reading about," infit MNSQ = 0.64, outfit MNSQ = 0.64). Item Wa5 ("Discussing my English writing with others is unenjoyable") misfit the L2 Writing Anxiety construct (infit MNSQ = 1.26, outfit MNSQ = 1.46). Two items misfit the Testing Anxiety construct, item Ta4 ("It seems to me that examination periods ought not to be made the tense situations which they are," infit MNSQ = 1.96, outfit MNSQ = 2.00) and item Ta8 ("I feel less confident during tests," infit MNSQ = 0.65, outfit MNSQ = 0.63).

For L2 Speaking Anxiety, the most difficult item and the easiest items to endorse were Sa7, "I'm afraid my partner will laugh when I speak English with a classmate in a pair" (Rasch item difficulty measure = 62.95) and Sa3, "I'm worried that my partner speaks better English than I do" (Rasch item difficulty measure = 42.21), respectively. These were also the most difficult and the easiest to endorse items for the entire questionnaire. Other "most difficult" and "easiest" items in terms of endorsability level were as follows: L2 Listening Anxiety most difficult, La6, "I get nervous and confused when I don't understand every word in listening test situations" (Rasch item difficulty measure = 59.13), easiest, La 11, "The thought that I may be missing key words frightens me" (Rasch item difficulty measure = 45.40); L2 Reading Anxiety most difficult, Ra9, "I usually end up translating word by word when I'm reading in English" (Rasch item difficulty measure = 54.71), easiest Ra2, "When reading English, I often understand the words but still can't understand what the author is saying" (Rasch item difficulty measure = 42.63); L2 Writing Anxiety most difficult, Wa10, "While writing in English, I'm nervous" (Rasch item difficulty measure = 56.38), easiest, Wa 8, "I worry that my English compositions are a lot worse than others'" (Rasch item difficulty measure = 44.85); Testing Anxiety most difficult, Ta7, "I feel my heart beating very fast during tests" (Rasch item difficulty measure = 55.49), easiest, Ta2, "I wish examinations did not bother me so much" (Rasch item difficulty measure = 40.79).

Table 1 Initial rasch item fit statistics for the five anxiety constructs

Item	Measure	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PMC
L2 speaking anxiety						
Sa1	41.58	**1.81**	8.0	**1.90**	7.7	0.59
Sa2	42.99	1.10	1.3	1.30	3.0	0.75
Sa3	42.21	1.17	2.0	1.20	2.1	0.71
Sa7	62.95	0.91	−1.2	0.85	−1.6	0.77
Sa10	59.70	0.90	−1.2	0.87	−1.5	0.78
Sa5	48.04	0.88	−1.4	0.89	−1.3	0.79
Sa11	46.31	0.88	−1.5	0.84	−1.9	0.81
Sa9	54.41	0.84	−2.0	0.83	−2.0	0.81
Sa8	54.12	0.81	−2.4	0.81	−2.2	0.81
Sa4	52.39	0.81	−2.5	0.80	−2.4	0.82
Sa6	45.30	0.77	−3.0	0.76	−2.9	0.81
L2 listening anxiety						
La1	56.63	**1.32**	3.6	**1.40**	4.2	0.65
La2	46.94	1.23	2.5	1.15	1.7	0.73
La5	46.29	1.14	1.6	1.16	1.8	0.70
La3	42.80	1.09	1.0	1.00	0	0.74
La7	54.23	1.02	0.3	1.05	0.6	0.73
La4	47.23	0.99	−0.1	0.98	−0.2	0.76
La11	45.40	0.90	−1.2	0.88	−1.4	0.77
La8	48.91	0.85	−1.8	0.89	−1.3	0.78
La6	59.13	0.87	−1.7	0.88	−1.5	0.78
La9	52.49	0.83	−2.1	0.85	−1.5	0.79
La10	49.93	0.76	−3.0	0.75	−3.1	0.81
L2 reading anxiety						
Ra8	46.82	**1.88**	8.6	**1.85**	8.4	0.51
Ra9	54.71	1.19	2.2	1.18	2.1	0.68
Ra2	42.63	1.08	0.9	1.08	1.0	0.63
Ra1	44.55	1.06	0.7	1.06	0.8	0.71
Ra4	50.94	1.00	0	0.98	−0.3	0.74
Ra6	47.81	0.97	−0.4	0.96	−0.4	0.72
Ra11	53.64	0.88	−1.5	0.88	−1.5	0.70
Ra5	49.26	0.86	−1.8	0.85	−1.9	0.75
Ra3	53.96	0.71	−3.9	0.72	−3.8	0.77
Ra7	53.40	0.71	−4.0	0.71	−3.9	0.80
Ra10	52.29	**0.64**	−5.2	**0.64**	−5.0	0.75
L2 writing anxiety						
Wa5	51.47	1.26	3.0	**1.46**	4.8	0.70
Wa2	52.14	1.14	1.7	1.26	2.9	0.72
Wa10	56.38	1.19	2.2	1.21	2.4	0.72

(continued)

Table 1 (continued)

Item	Measure	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PMC
Wa4	51.59	1.03	0.4	1.15	1.7	0.75
Wa1	46.75	0.98	−0.2	0.98	−0.2	0.76
Wa8	44.85	0.91	−1.1	0.90	−1.1	0.78
Wa3	47.30	0.89	−1.4	0.89	−1.3	0.79
Wa9	55.89	0.89	−1.3	0.89	−1.3	0.76
Wa6	47.66	0.88	−1.5	0.88	−1.4	0.77
Wa7	45.96	0.72	−3.8	0.70	−3.8	0.81
Testing anxiety						
Ta4	48.68	**1.96**	9.1	**2.00**	9.2	0.45
Ta1	48.56	1.12	1.4	1.21	2.4	0.66
Ta2	40.79	1.15	1.7	1.15	1.5	0.64
Ta3	49.07	0.92	−1.0	0.96	−0.4	0.71
Ta5	49.97	0.75	−3.3	0.75	−3.3	0.77
Ta7	55.49	0.73	−3.7	0.73	−3.7	0.75
Ta6	55.02	0.71	−3.9	0.72	−3.8	0.76
Ta8	52.42	**0.65**	−4.9	**0.63**	−5.1	0.78

Notes Bolded numerals indicate misfit (−0.7 to 1.3 MNSQ criteria). sa = Speaking Anxiety; la = Listening Anxiety; wa = Writing Anxiety; ra = Reading Anxiety; ta = Testing Anxiety; $N = 298$

Rasch Item and Person Reliability and Separation

Before removing the misfitting items, Rasch reliability and separation were obtained for the original 51 items by inputting items from each construct separately (Table 2). Rasch item reliability and separation were lowest for L2 writing anxiety (0.96, 4.72) and highest for L2 speaking anxiety (0.99, 9.39). Rasch person reliability and separation were lowest for Testing Anxiety (0.81, 2.07) and highest for L2 speaking anxiety (0.91, 3.27). In general, the five anxiety constructs showed reasonable person reliability. The person separation for four of the five constructs was below 3.00, indicating the ability of these instruments to separate participants

Table 2 Rasch reliability and separation estimates among the anxiety constructs

Type of anxiety	*k*	Item		Person	
		Reliability	Separation	Reliability	Separation
L2 speaking	11	0.99	9.39	0.91	3.27
L2 listening	11	0.97	5.84	0.89	2.79
L2 writing	10	0.96	4.72	0.87	2.57
L2 reading	11	0.97	5.47	0.86	2.52
Testing	8	0.98	6.32	0.81	2.07

Table 3 Rasch reliability and separation estimates excluding misfitting items

Type of anxiety	*k*	Item		Person	
		Reliability	Separation	Reliability	Separation
L2 speaking	10	0.99	9.71	0.92	3.36
L2 listening	10	0.97	5.55	0.88	2.71
L2 writing	9	0.96	5.09	0.86	2.46
L2 reading	9	0.97	6.23	0.86	2.48
Testing	6	0.98	6.70	0.77	1.83

into 3–4 groups. However, the person separation of Testing Anxiety was considerably lower than the L2 skills-related instruments and not ideal for an instrument related directly to testing. As a comparison, all misfitting items were removed from their respective constructs and the Rasch reliability and separation statistics were computed again (Table 3). Except for L2 Speaking Anxiety, all other constructs had lower person reliability and separation.

Rasch item-person maps were also obtained at this time to provide visual confirmation of the item hierarchy in each construct. The item-person maps used all items, prior to removing misfitting items (Figs. 3, 4 and 5). L2 speaking anxiety and L2 reading items were generally targeted appropriately for the sample population, while the mean endorsability level of L2 listening anxiety and L2 writing anxiety items were one standard deviation below the mean anxiety level of the participants. The endorsability level of testing anxiety items were almost two standard deviations below the anxiety level of participants. All five constructs showed some redundancy among items. L2 writing anxiety having five overlapping items and the least item spread, and L2 speaking anxiety having the greatest item spread.

Rasch Principal Components Analysis of Residuals

Misfitting items were kept for the initial Rasch Principal Components analysis of residuals (PCA or PCAR). Each individual construct was input separately to check construct dimensionality (Table 4). Rasch PCAR indicated support for unidimensionality across all five separate constructs, as the Rasch model explained well above 50 % for each construct and less than 10 % on principal contrasts. L2 Speaking Anxiety was the strongest construct (eigenvalue = 52.2, variance accounted for = 82.6 %), and L2 Reading Anxiety (eigenvalue = 17.3, varianced accounted for = 61.1 %) and Testing Anxiety were the weakest (eigenvalue = 12.6, varianced accounted for = 61.2 %). The eigenvalue was at or slightly above 2.0 for the principal contrast on three constructs (L2 Speaking Anxiety, L2 Listening Anxiety, and L2 Writing Anxiety); however, no contrasts explained more than 10 % variance for any construct. The principal contrast for Testing Anxiety explained 8.9 % of the variance, and the variance accounted for by the principal

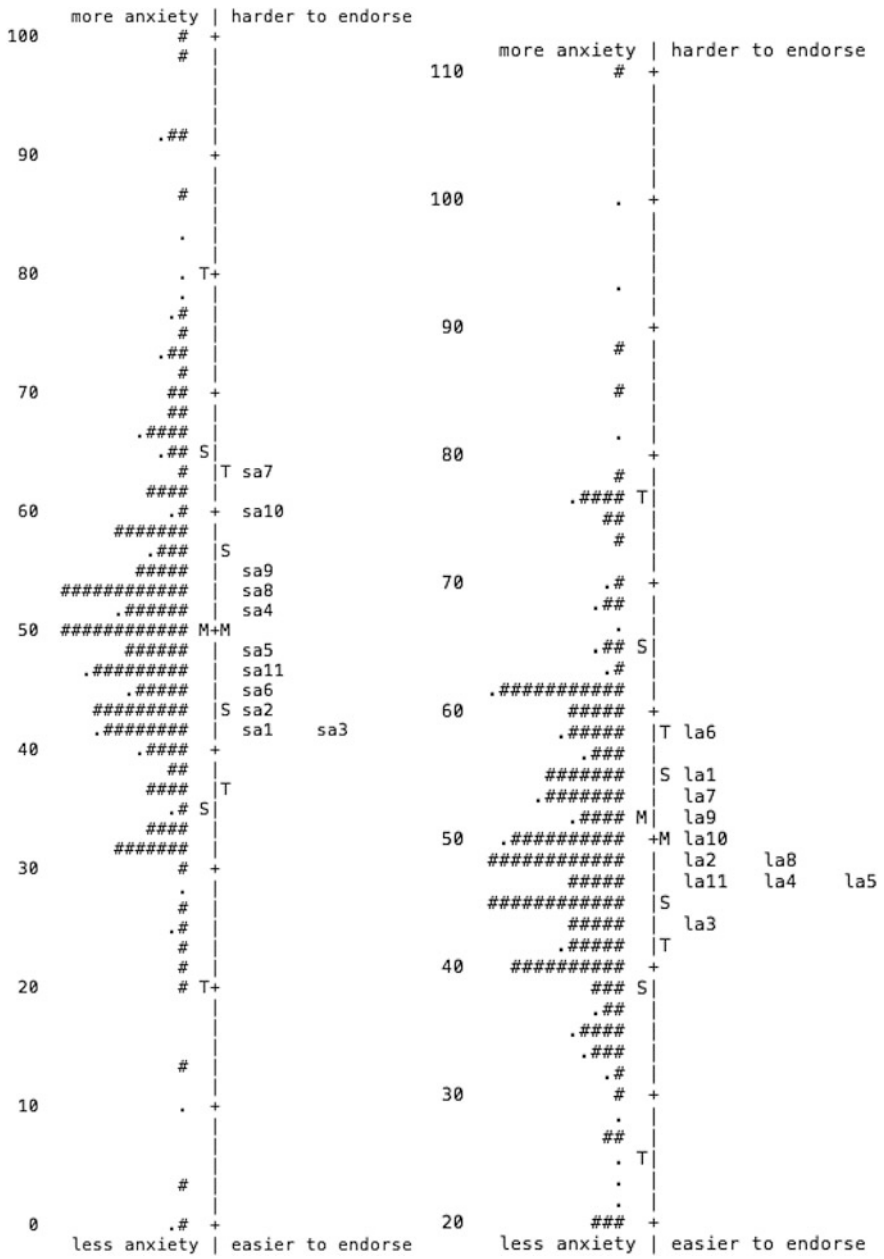


Fig. 3 The item-person maps of L2 speaking anxiety (*left*) and L2 listening anxiety (*right*). Each # is two people. $N = 298$

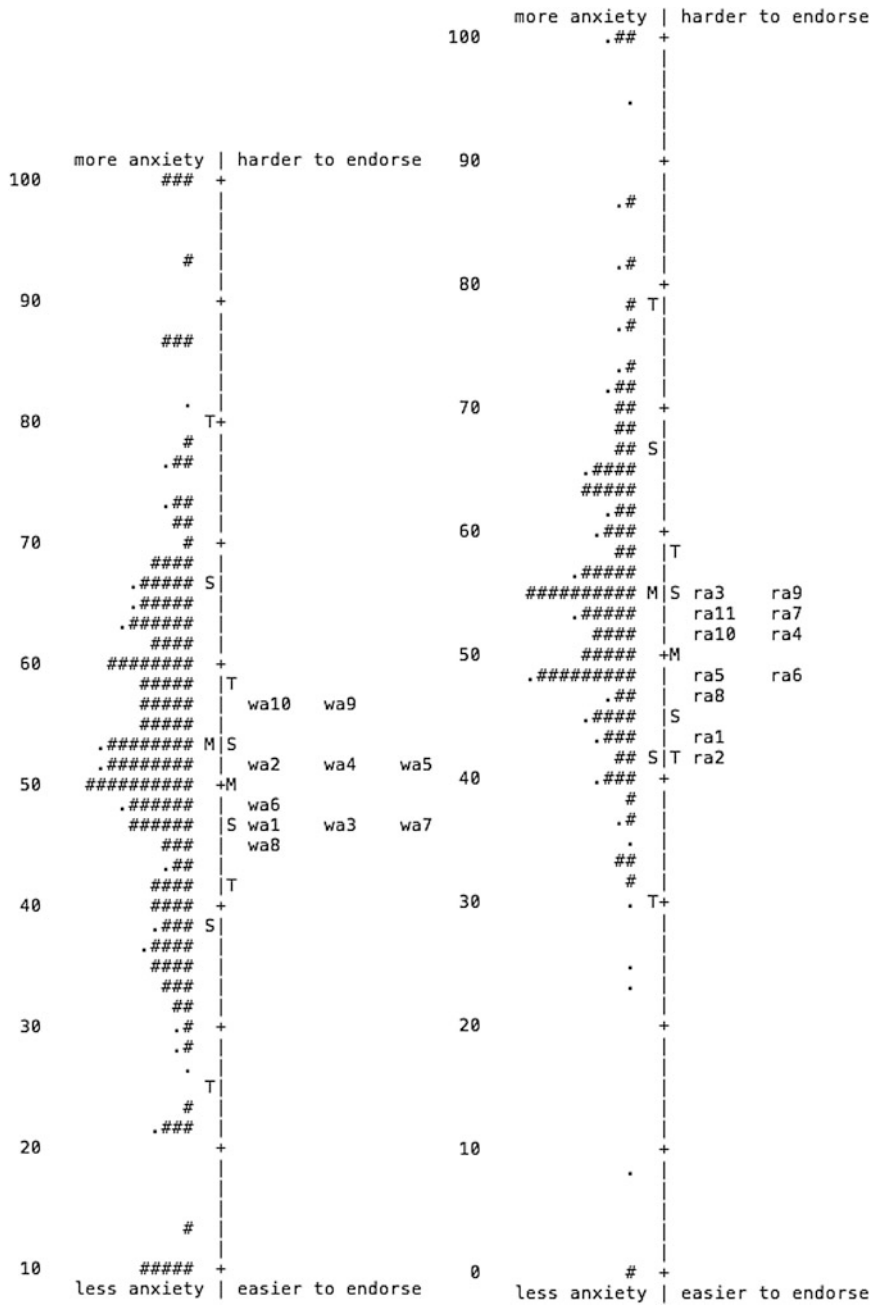


Fig. 4 The item-person maps of L2 Writing Anxiety (*left*) and L2 Reading Anxiety (*right*). Each # is two people for L2 Writing Anxiety and three people for L2 Reading Anxiety. $N = 298$

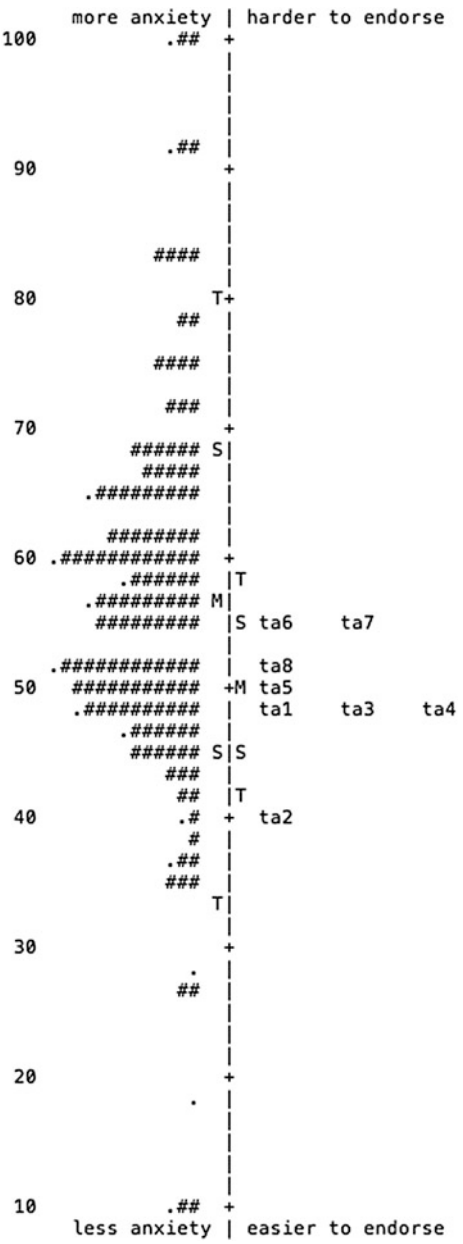


Fig. 5 The item-person map of Testing Anxiety. Each # is two people. $N = 298$

contrasts to L2 Listening Anxiety, L2 Writing Anxiety, and L2 Reading Anxiety were above 5.0, a possible indication of measurement noise.

As with Rasch reliability and separation analysis, all seven misfitting items were removed and Rasch PCAR was conducted again on each construct for a comparison

Table 4 Rasch PCA eigenvalues and variance accounted for (k = 51)

Type of anxiety	Rasch model		Principal contrast	
	Eigenvalue	% of var	Eigenvalue	% of var
L2 speaking	52.2	82.6	2.4	3.8
L2 listening	25.5	69.9	2.0	5.4
L2 writing	20.3	67.0	2.1	6.9
L2 reading	17.3	61.1	1.7	6.1
Testing	12.6	61.2	1.9	9.3

Table 5 Rasch PCA eigenvalues and variance accounted for, excluding misfitting items (k = 44)

Type of anxiety	Rasch model		Principal contrast	
	Eigenvalue	% of var	Eigenvalue	% of var
L2 speaking	58.1	85.3	2.1	3.0
L2 listening	21.7	68.4	1.9	5.9
L2 writing	18.0	66.7	2.2	8.0
L2 reading	20.9	69.9	1.5	5.0
Testing	11.5	65.8	1.9	10.9

(Table 5). The eigenvalue and variance accounted for by the Rasch model was higher in L2 Speaking Anxiety and L2 Reading Anxiety but lower in L2 Listening Anxiety, L2 Writing Anxiety, and Testing Anxiety.

Although Rasch model analysis is designed for individual constructs, Rasch PCAR can also provide some indication of relatedness among constructs with a single dataset by examining whether the contrasting dimensions to the Rasch model cohere into separate dimensions or comprise random noise. Therefore, as an initial test of relatedness among the five anxiety instruments, data from all five constructs were input into the Rasch model simultaneously. L2 Speaking and L2 Writing Anxiety items all loaded onto the primary, "positive" dimension. L2 Listening Anxiety, L2 Reading Anxiety, and Testing anxiety items all loaded onto the secondary, "negative" dimension (Fig. 6).

As a final examination of relations among the five anxiety constructs, a correlational analysis was conducted using Rasch person measures. First, correlations were computed for the constructs before removing misfitting items (Table 6). All five constructs were significantly correlated ($p < 0.05$). The strongest correlation was between L2 Listening Anxiety and L2 Reading Anxiety ($r = 0.71$), and the weakest correlation was between L2 Speaking Anxiety and Testing Anxiety ($r = 0.41$). After removing the seven misfitting items, the person measures were recalculated and the correlational analysis was conducted again (Table 7). Overall there were no appreciable differences in the Pearson's coefficient r values.

CON-TRAST	LOADING	INFIT MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM
1	.62	56.89	1.04	1.03	A 10	sa10
1	.62	53.17	1.10	1.14	B 9	sa9
1	.59	52.97	1.07	1.15	C 8	sa8
1	.58	51.75	.98	.98	D 4	sa4
1	.58	59.17	1.09	1.12	E 7	sa7
1	.56	47.43	1.08	1.06	F 11	sa11
1	.47	46.71	.91	.91	G 6	sa6
1	.43	45.03	1.21	1.19	H 2	sa2
1	.43	48.66	.91	.90	I 5	sa5
1	.40	44.47	1.33	1.46	J 3	sa3
1	.39	50.77	.79	.78	K 40	wa7
1	.35	50.33	.82	.81	L 41	wa8
1	.31	56.65	.88	.90	M 42	wa9
1	.30	51.38	.86	.86	N 34	wa1
1	.29	52.09	.84	.84	O 39	wa6
1	.28	44.01	1.68	1.81	P 1	sa1
1	.18	54.19	1.06	1.18	Q 38	wa5
1	.16	54.59	1.11	1.15	R 35	wa2
1	.08	51.72	1.06	1.05	S 36	wa3
1	.05	57.06	.89	.87	T 43	wa10
1	.01	54.25	1.01	1.01	U 37	wa4

LOADING	INFIT MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM
-.49	48.66	.87	.86	a 18	la7
-.48	45.35	.81	.79	b 19	la8
-.44	47.57	.80	.78	c 20	la9
-.41	43.76	1.02	1.00	d 16	la5
-.38	51.72	1.05	1.04	e 26	ra4
-.38	53.51	.84	.84	f 29	ra7
-.37	50.50	.79	.79	g 27	ra5
-.36	45.98	.73	.71	h 21	la10
-.35	49.45	.91	.90	i 28	ra6
-.33	44.15	1.11	1.18	j 13	la2
-.32	43.22	.87	.85	k 22	la11
-.32	53.07	.85	.85	l 49	ta6
-.32	51.82	.79	.80	m 17	la6
-.31	53.51	.84	.84	n 50	ta7
-.29	53.92	.82	.83	o 25	ra3
-.29	49.86	1.04	1.03	p 48	ta5
-.28	51.45	.96	.96	q 51	ta8
-.23	43.76	1.45	1.46	r 45	ta2
-.23	41.66	.97	.98	s 14	la3
-.22	48.73	1.56	1.58	t 30	ra8
-.21	44.33	.88	.86	u 15	la4
-.21	53.68	.78	.78	v 33	ra11
-.20	47.09	.76	.74	w 23	ra1
-.19	45.70	.81	.80	x 24	ra2
-.14	50.20	.98	1.01	y 12	la1
-.13	49.11	1.29	1.31	Z 46	ta3
-.10	48.90	1.09	1.11	Y 44	ta1
-.09	52.70	.62	.62	X 32	ra10
-.08	54.46	1.19	1.28	W 31	ra9
-.06	48.94	1.69	1.87	V 47	ta4

Fig. 6 The Rasch PCA of residuals for all items (k = 51). sa = Speaking Anxiety; la = Listening Anxiety; wa = Writing Anxiety; ra = Reading Anxiety; ta = Testing Anxiety

Table 6 Correlations among the five anxiety constructs (k = 51)

	SA	LA	WA	RA	TA
L2 speaking anxiety	–	0.55	0.56	0.47	0.41
L2 listening anxiety	.	–	0.51	0.71	0.51
L2 writing anxiety			–	0.69	0.58
L2 reading anxiety				–	0.57
Testing anxiety					–

Note Correlations were all significant ($p < 0.01$); N = 298

Table 7 Correlations among the five anxiety constructs excluding misfitting items (k = 44)

	SA	LA	WA	RA	TA
L2 speaking anxiety	–	0.54	0.58	0.50	0.39
L2 listening anxiety	.	–	0.50	0.70	0.52
L2 writing anxiety			–	0.67	0.56
L2 reading anxiety				–	0.56
Testing anxiety					–

Note Correlations were all significant ($p < 0.01$); N = 298

Discussion

While the top performing items of existing L2 anxiety instruments overall function well, the Rasch item fit analysis demonstrates that there is still room for improvement. The L2 Speaking Anxiety item (Sa1) that misfit the construct seemed more related to perceptions of speaking competence rather than anxiety ("…other students speak better than I do."). The L2 Listening Anxiety item (La1) that misfit the construct contained vague wording ("I get stuck with…") that could apply to other types of anxiety, or be related to comprehension rather than anxiety. Of the two L2 Reading Anxiety item that misfit, one (Ra4) was more related to speaking than to reading ("…that I can't pronounce), while the other (Ra10) was, to be blunt, extremely silly. The original item was written for Japanese learners just beginning to learn how to read English; the item was likely intended to tap into the orthographic challenge of reading alphabetic characters in contrast to Chinese ideograms (*kanji*) and Japanese syllabary characters (*kana*). However, the wording "funny letters and symbols" has little if anything to do with reading anxiety, and the Rasch item analysis bears this out.

The L2 Writing Anxiety had only one misfitting item (Wa5), which could be construed as a speaking item as well as writing ("Discussing my writing…"). Additionally, there is a question whether asking if discussing writing is "enjoyable" constitutes language anxiety or not. Finally, there were two items that misfit the Testing Anxiety construct. The first (Ta4) contained awkward wording, including a negation and a "seems to me" clause, neither of which gives the sense of anxiety. The second (Ta8) was about confidence rather than anxiety. Feeling "less confident" does not necessarily imply "anxiety"—as discussed in the background section, conceptually there are still wording issues with many anxiety instruments, as anxiety is seen in clinical psychology as abnormal and not simply being "nervous" or "concerned." Since this construct had the lowest variance accounted for by the Rasch model as well as the lowest Rasch person reliability it is worth noting that the study from which these items were borrowed had an extremely low sample size ($N = 79$) to go with a large number of items ($k = 57$) which originally stemmed from two separate studies (Fujii 1993; Sarason 1975). The "test anxiety" instrument created used traditional factor analysis to reduce the 57 items to 12 items across three factors, one of which was related to mental and psychical exhaustion. The results in the present study indicate problems with item fit and construct validity, leading to the logical conclusion that this construct still needs seriously reworking before further usage.

While Rasch model analysis is designed for single constructs rather than a multi-construct questionnaire like the one used in this study, the Rasch PCAR was revealing concerning the relation among the five anxiety constructs. When all items were input into the Rasch model, the Rasch model correctly identified all L2 Speaking Anxiety items as a single construct; all SA items had "positive" loadings above 0.40. However, all the L2 Writing Anxiety items also loaded onto the primary dimension, an indication that study participants treated speaking and writing

items similarly; this may not be surprising, as speaking and writing are considered "active," student-output related L2 skills. The Rasch PCAR also indicated the relationship between L2 Reading Anxiety and L2 Listening Anxiety; all the RA and LA items loaded "negatively" onto the secondary dimension. This result may be the result of study participants treating reading and listening as "passive," input-related L2 skills. Testing Anxiety (TA) items also loaded negatively. This may not be too surprising given the context of the study: in Japan, university EFL programs often use standard exams such as the Test of English for International Communication (TOEIC), which is used by Japanese companies for hiring and internal promotion. While there is a version of the exam that tests speaking and writing, the TOEIC used in universities (TOEIC IP) typically only tests reading and listening abilities. Thus, it makes sense that Rasch identified L2 listening and reading anxieties as related strongly to test anxiety.

The correlational analysis of Rasch person measures among the five anxiety instruments contradicts previous assumptions in the L2 anxiety literature that test anxiety is not related to L2 anxiety. The results in this study indicate that the various L2 anxieties are, in fact, strongly related to each other as well as to testing anxiety. Indeed, this should come as no surprise: the wording of the L2 listening, writing, and reading items tend to focus on exam and assessment situations. In particular, nearly all L2 writing anxiety items are related to concerns about written essay assessment (and hence, testing). Given the relation of all four L2 anxieties to testing anxiety—which is supposed to be the same regardless of first or second language context—the results additionally suggest that the use of a second language may not represent a separate anxiety, agreeing with Dewaele (2013) and Jung and McCroskey (2004).

Conclusion

The results of this study provide examples of a construct created through Rasch model principles (L2 Speaking Anxiety) compared to constructs created through correlation and factor analysis (L2 Listening Anxiety, L2 Reading Anxiety, L2 Writing Anxiety, Testing Anxiety). Whereas a measurement instrument created according to Rasch principles assumes that questionnaire takers with a high level of anxiety will have a higher probability of endorsing items that also indicate a higher degree of anxiety (Wilson 2005), a measurement instrument created through traditional statistics assumes that someone with a greater level of anxiety will endorse all items to a greater degree than someone with a lower level of anxiety. As Waugh and Chapman (2005) have pointed out, the fact that items correlate strongly in traditional factor analysis does not indicate construct validity. The use of multiple items to artificially inflate Cronbach's alpha, rather than spread participants across a range of endorsability levels on the hypothesized construct, also does not lead to greater construct validity.

A limitation of the current study is that all study participants were in a single EFL context (Japan), as well as roughly the same age and level of education.

The sample size ($N = 298$) was adequate, but given the number of total items ($k = 51$) a larger sample size may be desirable. Finally, the number of Likert-type categories (six) had to be collapsed for the four of the five measurement instruments that were originally constructed using traditional statistical methods. This may have slightly suppressed the person reliabilities and separation for these constructs.

Whether the "four L2 anxieties" truly measure separate anxieties may still be open for debate. Certainly several items in each anxiety measurement, and particularly those of testing and L2 reading anxiety, are in need of revision before further investigation of the relationship among the varying hypothesized aspects of L2 anxiety. Deeper analysis of the relationship among the "active" and "passive" anxieties is also warranted, once individual "four anxieties" L2-skills based measurement instruments have been further examined and validated in other ESL or EFL contexts.

Notes

1. The full 51-item questionnaire was omitted for space considerations but is available by request; please email the author.
2. The original L2 Reading Anxiety instrument was created by Saito et al. (1999), but as only descriptive statistics were used in that paper, Matsuda and Gobel's factor loadings were used to select items.
3. Likert category utility statistics were omitted for space considerations but are available by request.

References

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38*, 665–680.

Apple, M. T. (2013). Using Rasch analysis to create and evaluate a measurement instrument for foreign language classroom speaking anxiety. *JALT Journal, 35*(1), 5–28.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum.

Cheng, Y., Horwitz, E. K., & Schallert, D. L. (1999). Language anxiety: Differentiating writing and speaking components. *Language Learning, 49*, 417–446.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Dewaele, J.-M. (2013). The link between foreign language classroom anxiety and psychoticism, extraversion, and neuroticism among adult bi- and multilinguals. *Modern Language Journal, 97*(3), 670–684.

Fujii, Y. (1993). Tesuto eikyou inbentorii (TII) no sakusei [Construction of a Test Influence Inventory (TII)]. *Japanese Journal of Psychology, 64*(2), 135–139.

Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. London: Edward Arnold.

Green, S. B., Lissitz, R. W., & Mulait, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827–838.

Horwitz, E. K., Horwitz, M. B., & Cope, J. A. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*, 125–132.

In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System, 34,* 317–340.

Jung, H.-Y., & McCroskey, J. C. (2004). Communication apprehension in a first language as predictors of communication apprehension in a second language: A study of speakers of English as a Second Language. *Communication Quarterly, 52*(2), 170–181.

Kaipper, M. B., Chachamovich, E., Hidalgo, M. P. L., Torres, I. L. S., & Caumo, W. (2010). Evaluation of the structure of Brazilian State-Trait Anxiety Inventory using a Rasch psychometric approach. *Journal of Psychosomatic Research, 68*, 223–233.

Kim, J.-H. (2000). Foreign language listening anxiety: A study of Korean students learning English. (Unpublished doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 30004305).

Kimura, H. (2008). Foreign language listening anxiety: Its dimensionality and group differences. *JALT Journal, 30*(2), 173–195.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85–106.

Linacre, J. M. (2007). *A user's guide to WINSTEPS: Rasch-model computer program*. Chicago: MESA.

MacIntyre, P. D., & Gardner, R. C. (1989). Anxiety and second language learning: Toward a theoretical clarification. *Language Learning, 39*, 251–275.

MacIntyre, P. D., & Gardner, R. C. (1991). Language anxiety: Its relationship to other anxieties and to processing in native and second language. *Language Learning, 41*, 513–534.

Matsuda, S., & Gobel, P. (2004). Anxiety and predictors of performance in the foreign language classroom. *System, 32*(1), 21–36.

Price, M. L. (1991). The subjective experience of foreign language anxiety: Interviews with highly anxious students. In E. K. Horwitz & D. J. Young (Eds.), *Language anxiety: From theory and research to classroom implications* (pp. 101–108). Englewood Cliffs, NJ: Prentice-Hall.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago.

Saito, Y., Horwitz, E. K., & Garza, T. J. (1999). Foreign language reading anxiety. *Modern Language Journal, 83*, 202–218.

Sarason, I. G. (1975). The Test Anxiety Scale: Concept and research. In I. G. Sarason & C. D. Spielberger (Eds.), *Stress and anxiety* (Vol. 2, pp. 193–217). Washington, D.C.: Hemisphere.

Spielberger, C. D. (1983). *Manual for the state-trait anxiety inventory*. Palo Alto, CA: Consulting Psychologists Press.

Tenenbaum, G., Furst, D., & Weingarten, G. (1985). A statistical reevaluation of the STAI anxiety questionnaire. *Journal of Clinical Psychology, 41*(2), 239–244.

Waugh, R. F., & Chapman, E. S. (2005). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement, 6*(1), 80–99.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wolfe, E. W., & Smith, E. V, Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I—Instrument development tools. *Journal of Applied Measurement, 8*(1), 97–123.

Wright, B. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3*(1), 3–24.

Young, D. J. (1999). *Affect in foreign language and second language learning: A practical guide to creating a low-anxiety classroom atmosphere*. Boston: McGraw Hill.

Author Biography

Matthew T. Apple (MFA, MEd, EdD) is Associate Professor in the Department of Communication, Ritsumeikan University, Kyoto, Japan. Recent publications include *Language Learning Motivation in Japan* (2013, Multilingual Matters) and articles in *JALT Journal* and the *Journal of Applied Measurement*. His research interests include ESP, individual differences, and educational statistics.

Examining the Psychometric Quality of a Modified Perceived Authenticity in Writing Scale with Rasch Measurement Theory

Nadia Behizadeh and George Engelhard Jr.

High-stakes, large-scale testing has proliferated in the United States, and a plethora of studies indicate that instructional practices have suffered (e.g., Darling-Hammond 2010). In particular, researchers theorize that although US students are writing more (Applebee and Langer 2011), classroom writing experiences are not highly authentic to students, especially for urban, students of color who are economically disadvantaged (Ball and Ellis 2008). Authenticity is a key motivational variable in school settings, and if US students are not experiencing authentic writing instruction, then reasons for this lack, such as the potential negative effects of current large-scale writing assessments, need to be addressed. Alternately, if some students are experiencing highly authentic writing instruction, a closer examination of factors that contribute to the enactment of authentic writing instruction needs to occur. However, because authenticity is a student's perception of the meaningfulness of instruction, student perspectives are needed to explore the authenticity of writing instruction in the United States.

There are not many tools available for measuring authenticity. One past tool is the Perceived Authenticity in Writing (PAW) Scale designed to measure perceived authenticity in writing instruction for adolescents for a *specific task* (Behizadeh and Engelhard 2014). However, there is a need for a similar scale to the PAW Scale, but one that could be used for examining students' general impression of their writing instruction as a whole. This would allow researchers, educators, and policymakers

N. Behizadeh (✉)
Department of Middle and Secondary Education, College of Education and Human Development, Georgia State University, P.O. Box 3978, Atlanta, GA 30302-3978, USA
e-mail: nbehizadeh@gsu.edu

G. Engelhard Jr.
Educational Psychology, The University of Georgia, Aderhold Hall, Athens, GA 30602, USA
e-mail: gengelh@uga.edu

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_5

to administer the instrument across a multitude of contexts, and then (1) compare perceptions of authenticity; (2) identify schools or districts with high or low authenticity for deeper qualitative examination; and (3) analyze correlations among authenticity and other variables, such as socioeconomic status. In particular, this last purpose for a general authentic writing scale would allow researchers to identify differential access to authentic writing instruction and to explore issues related to social justice in writing assessment.

To this end, a modified instrument was created: the Modified Perceived Authenticity in Writing (MPAW) Scale. The MPAW Scale asks students to evaluate their overall impression of the authenticity of the current writing instruction that they are receiving. This chapter examines the psychometric properties of the MPAW Scale for use in a larger study of perceived authenticity among urban students of color in the US. The following research questions guided this study:

1. Does the internal structure of the MPAW Scale represent gradations of item difficulty?
2. Do the MPAW items exhibit acceptable model-data fit that supports the validity of inferences regarding student perceptions of the authenticity of their writing instruction?
3. Do the MPAW items exhibit measurement invariance when explored with explanatory variables such as grade level, gender, and student attitude toward writing?

Literature Review

What Is Authenticity?

Although numerous scholars call for increasing the authenticity of literacy education and writing education in particular, the meaning of "authentic" is somewhat unclear. In past research, educational authenticity has traditionally been defined as the connection of a school task to the real world (Newmann et al. 1996; Purcell-Gates et al. 2012; Seunarinesingh 2010). However, drawing on the idea of authenticity as subjective (Ashton 2010; Splitter 2009), past research (Behizadeh 2014, 2015) has presented a definition of authenticity in writing as a student's *perception* that a writing task connects to their life. This perception of authenticity includes culture, personal interests, and community or global issues that matter to the student. Importantly, this definition establishes that the authority for determining authenticity resides in the student, not with teachers or policymakers. Educators and researchers may hypothesize that particular tasks are highly authentic for students, but without confirmation from students that these tasks are indeed meaningful and relevant to their lives, a strong claim for authenticity cannot be made.

Why Does Authenticity in Writing Matter?

A large body of research supports authenticity in education as a key component for increasing student engagement and achievement, particularly in teaching writing (Fisher 2007; Freire 1970/2000; Morrell 2008; Purcell-Gates et al. 2007; Sisserson et al. 2002; Winn and Johnson 2011). In Hillocks' (2011) review of a century of literacy research, he stated, "We know from a very wide variety of studies in English and out of it, that students who are authentically engaged with the tasks of their learning are likely to learn much more than those who are not" (p. 189). Across literacy research connected to authenticity, there is the belief that greater authenticity increases student engagement and achievement. In essence, perceived authenticity by students can serve an important motivational role in educational settings. Additionally, standards for English language arts stress the importance of "authentic, open-ended learning experiences" (National Council of Teachers of English and International Reading Association 2012, p. 6) for student learning, as do documents outlining twenty-first century skills (Partnership for twenty-first Century Learning n.d). Additionally, the Common Core State Standards for English language arts (Common Core State Standards Initiative 2015) emphasize writing for real audiences and publishing and distributing student writing. Finally, connecting instructional content to the real world is a consideration of standards used to evaluate teachers (Council of Chief State School Officers 2011).

What Factors May Be Impeding Authenticity in Writing Instruction?

In addition to wide support for the importance of authenticity in writing instruction, literacy researchers have documented the misalignment between writing assessments that focus on conventions and mechanics and a definition of writing as an iterative, social, and creative contextualized process (Au and Gourd 2013; Dyson and Freedman 2003). According to leading assessment scholars, "The overreliance on psychometric approaches to assessment risks reducing diversity in teaching, learning, and assessment practices; dismissing alternative disciplinary experiences; and marginalizing local knowledge and expertise" (Haertel et al. 2008, p. 77). Thus, there is a conflict between high-stakes writing assessments that encourage rote writing instruction and research and standards supporting meaningful, authentic writing instruction.

One way to position an argument for examining the authenticity of writing instruction in relation to writing assessment is through a validity lens (Messick 1995; Kane 2013). According to Messick and Kane, validity is the use of a test for a particular purpose (in this case, to evaluate writing achievement), and validity is

evaluated by developing an argument that includes multiple sources of evidence. Messick (1995) offered a unified theory of validity, stating, "Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment" (p. 741). Condensing Messick's (1995) six connected aspects of validity to two, major sources of validity evidence are (1) the match between the theorized construct of the assessment (definition, processes, structural elements) and the representation of this construct in the assessment; and (2) the consequences of assessment. The degree to which assessment practices impact instruction, including positive or negative effects on perceived authenticity, are part of consequential validity. However, construct and consequential validity are related. Slomp et al. (2014) articulated that issues with construct validity, especially misrepresenting or under representing the construct of writing, are closely connected to issues with consequential validity. Applying this to authenticity, if assessments are based on the idea of writing as a decontextualized set of skills, then this construction of writing can lead to teaching practices that focus on building skills without attending to the epistemic and identity-related aspects of writing.

Because of the importance of consequential validity in evaluating a writing assessment, an instrument for measuring the perceived authenticity of writing instruction by students is a tool that can be a useful for collecting validity evidence. If particular assessments are impeding authentic writing instruction, these assessments may need to be revised based on data collected from instruments such as the MPAW Scale.

Is There Evidence that Large-Scale Writing Assessments Are Reducing Authenticity?

A wide range of research indicates that high-stakes, large-scale writing assessment is impeding authentic writing instruction. In a qualitative study, Luna and Turner (2001) interviewed teachers administering high-stakes writing tests in Massachusetts, and the authors reported that teachers felt they were teaching to the test instead of providing rich writing instruction. A focus on ensuring students learned the formula for the five-paragraph essay rather than effective communication was critiqued as a negative outcome of high-stakes writing tests. In another study conducted in North Carolina, researchers found that high-stakes writing assessment resulted in "form over content and product over process" (Watanabe 2007, cited in Au and Gourd 2013, p. 17). Similarly, in their review of writing research, Dyson and Freedman (2003) argued that quality of writing depends on students' investment in a topic and their need to communicate information,

constituting a validity problem for standardized writing tests that are not compelling to students. Furthermore, negative washback is more pronounced for culturally and linguistically diverse students (Ball and Ellis 2008; Madaus 1994). In Ball and Ellis' (2008) review of decades of writing research, they concluded "that students of color are disproportionately relegated to classrooms using drill exercises rather than interactive, meaningful approaches that require extended writing, reflection, and critical thinking" (p. 507). However, although researchers can hypothesize that instruction is not authentic to students, an instrument that collects students' judgments of authenticity could help examine the degree to which certain assessments are affecting authenticity.

Methods

Again, our research questions are: (1) Does the internal structure of the MPAW Scale represent gradations of item difficulty? (2) Do the MPAW items exhibit acceptable model-data fit that supports the validity of inferences regarding student perceptions of the authenticity of their writing instruction? and (3) Do the MPAW items exhibit measurement invariance when explored with explanatory variables such as grade level, gender, and student attitude toward writing? To answer these questions, our analytic approach relied on invariant measurement. Invariant measurement (Engelhard 2013) draws on Rasch measurement theory (Rasch 1960/1980), and this framework is used to investigate the psychometric quality of the MPAW Scale. Invariant measurement is based on the requirement that instruments, including their meaning and use, remain consistent across different subgroups of students. If we create a stable and invariant frame of reference, then we begin to consider substantive differences in student perspectives within and between students. The Facets computer program was used to produce Wright maps that visually depict the relationships among persons (students), items, and other variables, as well as model-data fit statistics that provide support for inferences regarding how well the Wright map represents the latent variable of perceived authenticity (Linacre 1989).

For research questions one and two regarding item difficulties and model-data fit, we created Model 1 that included only persons and items. For research question three regarding explanatory variables, we created Model 2 that included persons, items, and three explanatory variables: student attitude toward writing, gender, and grade level. Both models used the partial credit model and the equations for each model are presented below:

Model 1

$$\ln \frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \tau_{ik}$$

Model 2

$$\ln \frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \Delta_j - \tau_{ik}$$

where

P_{nijk} the probability of student n responding in category k on item i;
P_{nijk-1} the probability of student n responding in category $k - 1$ on item i;
θ_n the perception of authenticity by student n;
δ_i the location of item i;
Δ_j the location of explanatory variable j; and
τ_{ik} the difficulty of responding in category k relative to $k - 1$ on item i.

The explanatory variables, Δ_j, included in this study are grade, gender and attitude toward writing.

Instrument

The Modified Perceived Authenticity in Writing (MPAW) Scale consists of 16 items with a 6-point Likert Scale (1-Strongly Disagree to 6-Strongly Agree) (see Appendix for details). One original item from the PAW Scale was dropped, and language was changed from task-specific to general for all other items. For example, an original item on the PAW Scale states, "Writing this paper helped me to understand the topic better" and the corresponding item on the MPAW Scale states, "Writing in my English language arts class helps me to understand topics better." In addition to responding to the 16 items, students also indicated their grade level (6–9), their interest in writing (from 1–6 with 1 indicating the lowest interest and 6 representing the highest interest), and their gender (Male, Female, or Other.) These demographic questions were used to explore how the scale may function differently for different groups of students.

Participants

Seventy-four students at one school site completed the MPAW Scale during an after-school program in the spring of 2015. All students provided written assent, and they had written parental consent to participate. Students mostly identified as Black or African American, 99 % participated in free and reduced school lunch programs, and their ages ranged from 11- to 14-years old.

Data Collection and Data Analysis

The MPAW Scale was administered in paper format to the 74 students in the study by a research assistant. Demographic questions preceded the MPAW Scale items. The research assistant first introduced the purpose of the study, indicating that our goal was to understand students' views on their current writing instruction in their English language arts class. Then the research assistant read through the demographic questions and instructed students to choose the responses that best represented them. Next, the research assistant read through the instructions for the MPAW Scale and then read the items, pausing after each item so students could record their answers. There were also additional Likert items and two short answer questions students answered in addition to the MPAW Scale items, but these items are not analyzed in the current study.

After data were collected, all data were entered into an Excel spreadsheet and then exported to the Facets program to run the Rasch analyses, including conversion of the raw ordinal data into interval data, creation of person and item fit statistics, and creation of Wright maps that visually display person and item locations and any additional variables included in particular models. For our analyses, the original 6-point Likert scale structure was maintained and ratings were not collapsed. Our findings based on these analyses are described in detail in the next section.

Findings

Model 1 Results

The first analysis explores if the internal structure of the MPAW Scale represents gradations in item difficulty and if items exhibit acceptable model-data fit. Overall, the Rasch model explained 51.7 % of the variance of the MPAW Scale. The results of the Rasch model indicated that the MPAW Scale has reasonably high reliability of person separation with $Rel_{Students} = 0.89$. Additionally, the scale has a relatively high reliability of item separation statistic with $Rel_{Items} = 0.77$. Table 1 contains the summary statistics for Model 1. These relatively high reliability statistics indicate that the MPAW Scale's items can be separated from one another and can be used to differentiate students' perceptions of authenticity, suggesting that the internal structure of the scale does indeed represent gradations of item difficulty.

Figure 1 contains the Wright map for Model 1, which displays the spread of persons and items graphically. This figure visually presents the item and person

Table 1 Summary statistics for Model 1

	Students	Items
Mean	0.32	0.00
SD	1.33	0.22
N	74	16
Infit M	1.11	0.99
Infit SD	0.93	0.31
Outfit M	1.16	1.16
Outfit SD	1.17	0.67
Reliability of separation	0.89	0.77
Chi-square	606.0*	68.7*
Df	73	15

*$p < 0.01$

separation statistics noted above. Column 1 is the scale in logits that acts as a common *ruler* in Rasch measurement theory to examine the relationship between persons and items. The next two columns present the locations of persons and items on the logistic scale. Higher items on this scale indicate lower levels of endorsement, meaning the item was more "difficult" to endorse, whereas lower locations for items indicate higher levels of endorsement.

Thus, Item 6, "I discuss the topics of my ELA writing assignments with my family," was the most difficult item to endorse, and Item 11, "I am proud of what I write in my ELA class," was the easiest item to endorse. Based on the trend in the literature of school writing tending to be a classroom contained activity versus a broader activity connecting to the community, these levels of endorsement make theoretical sense.

The person separation statistics, item separation statistics, and Wright map indicate that there is a hierarchy of item difficulty. Moving to the second research question regarding model-data fit, Wright and Linacre (1994) suggest that acceptable indices of model-data fit are obtained when the Infit and Outfit statistics range from 0.60 to 1.40. Table 2 presents the item quality index for all items in the scale, and Infit and Outfit statistics. As can be seen in Table 2, there are five items exhibiting some misfit. Item 1 is the only item with both Infit and Outfit statistics outside the acceptable range, while Items 2, 3, and 4 had Outfit statistics outside the acceptable range. Item 12 had a generally higher than acceptable Infit statistic.

These misfitting items suggest that there may not be a consistent difficulty hierarchy for these items. This makes sense when thinking about authenticity as a subjective judgment because students may value particular factors of authenticity above others and additionally, students may perceive their writing instruction differently. We return to the misfitting items in the discussion section. We also propose modifications that may improve model-data fit.

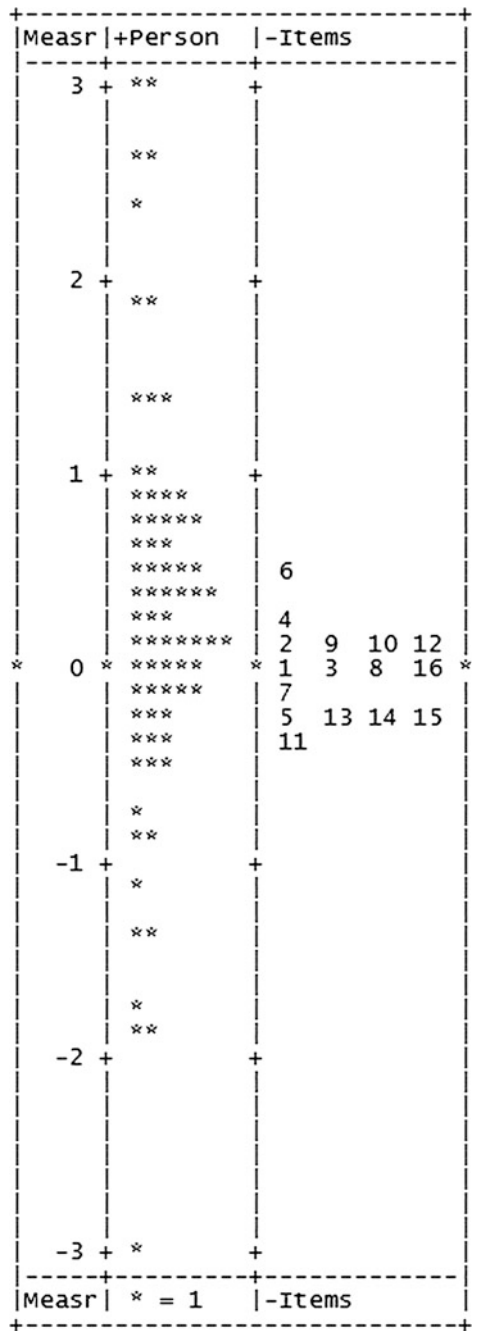


Fig. 1 Wright map for Model 1

Table 2 Item quality index in Rasch analysis

Items	Mean Rating	Measure	SE	Infit MS	Outfit MS
1. The writing that I do in my ELA class is related to my life outside of class	3.88	0.06	0.12	1.93*	2.38*
2. I enjoy writing in my ELA class	3.84	0.13	0.10	0.64	0.56*
3. ELA writing assignments relate to topics I care about in the world	3.85	0.05	0.10	1.14	2.29*
4. People other than my teacher read the papers I write for school	3.62	0.25	0.09	1.39	2.20*
5. I will use what I am learning about writing to write other papers in the future	4.24	−0.21	0.10	0.82	0.82
6. I discuss the topics of my ELA writing assignments with my family	3.42	0.48	0.10	0.88	0.85
7. What I am learning about writing is important for my life	4.08	−0.11	0.11	0.89	0.87
8. ELA assignments are important to me	4.00	0.00	0.1	0.82	0.74
9. Writing in my ELA class connects to my personal interests	3.79	0.17	0.10	0.92	1.12
10. People who read my ELA writing assignments will change their opinions, actions, or feelings	3.91	0.11	0.10	0.92	0.91
11. I am proud of what I write in my ELA class	4.36	−0.38	0.12	1.13	1.04
12. I discuss the topics of my writing assignments with friends	3.81	0.18	0.10	1.80*	1.04
13. I am gaining writing skills that I will use later in my life in my ELA class	4.34	−0.23	0.11	0.89	0.79
14. Writing in my ELA class helps me develop my thoughts, opinions, or beliefs	4.36	−0.25	0.12	0.78	0.81
15. Writing in my ELA class is making me a better writer	4.26	−0.24	0.11	0.73	0.78
16. Writing in my ELA class helps me to understand topics better	4.00	0.00	0.11	0.81	0.78

Notes (1) English language arts is abbreviated to ELA in this Table
(2) *Indicates Infit and Outfit statistics indicating item misfit

Model 2 Results

The second model included items and persons, and also added three explanatory variables. This model was used to answer the third research question: Do the MPAW items exhibit measurement invariance when explored with explanatory variables such as grade level, gender, and student attitude toward writing? Based on our analyses, differences on MPAW Scale by all explanatory factors were

statistically significant at a 0.01 level. This means that in this study, perceptions of authenticity varied significantly by grade level, by gender, and by student attitude toward writing. This finding aligns with past research (Behizadeh 2014, 2015) that found that authenticity varied by student characteristics, including gender and ethnicity. Potentially due to shared characteristics of a subgroup (e.g., cultural background), particular subgroups within a student population may perceive the authenticity of writing instruction at different levels.

Figure 2 is the Wright map for Model 2, and it graphically displays the relationships between items, students, and explanatory variables. As can be seen in this graphic representation, students who had high interest in writing (represented by a 6 on the 6-point scale) also were more likely to have higher scores on the MPAW Scale. This connection is theoretically logical since both authenticity and writing interest are motivational variables and a highly motivated student who has a high interest in writing may perceive instruction as more authentic and connected to their life. Additionally, males and those who indicated "Other" for gender were more likely to have higher scores on the MPAW Scale than females, a difference that does not have a clear explanation based on the literature. Also, in terms of grade level, those in the 6th grade were more likely to have higher scores on the MPAW Scale while those in the 8th grade were more likely to have lower scores. As a possible explanation for this difference, students in different grades are experiencing different curricula, which may include features that raise or lower perceptions of authenticity; yet this hypothesis cannot be confirmed by the data collected in the current study. Although there already exists a strong theoretical rationale for the connection between higher scores on the MPAW Scale and higher interest in writing, the underlying reasons for males and 6th graders having higher scores would need to be confirmed in future studies drawing on larger sample sizes and also explored further with qualitative methods.

Regarding interactions between items and explanatory variables, there were no significant interactions between items and student attitude toward writing, grade level, or gender. Thus, the MPAW items exhibit measurement invariance when explored with the explanatory variables of grade level, gender, and student attitude toward writing.

Discussion

Returning to research questions 1 and 2, the Wright map for Model 1 suggests a hierarchy of item difficulties, and there was high reliability of item and person separation. Regarding research question 3, the MPAW items exhibit measurement invariance when explored with the explanatory variables of grade level, gender, and student attitude toward writing. These findings suggest that the scale is able to differentiate students in terms of perceived authenticity, and that the internal structure of the MPAW is stable. Using the principles of invariant measurement, we identified several misfitting items. These items may not be consistently interpreted

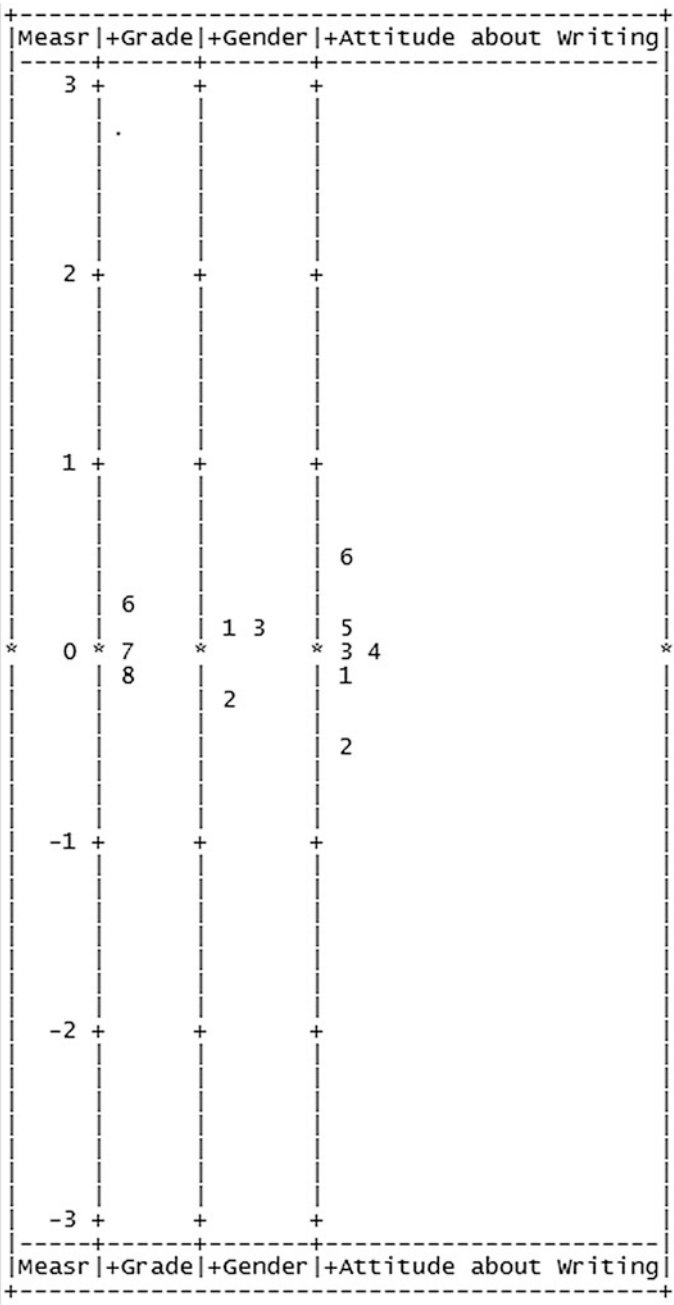


Fig. 2 Variable map for Model 2 [*Notes* Gender (1 = Male; 2 = Female; 3 = Other); Attitude (1 = low interest; 6 = high interest)]

in terms of a hierarchal structure for all items. However, unlike with the measurement of achievement in certain content areas (such as math) where there should be a more or less orderly progression of difficulty based on students levels of knowledge, the measurement of affective variables is more subjective; it may be that some students value certain features of authenticity more than other students. Additionally, because the questions are around how students perceive authenticity, individual students may perceive the same writing instruction differently.

However, revisions to the MPAW Scale could result in better model-data fit. For example, looking at the data, we found that students are not using all categories; the lower end of the scale was underutilized in this study, and other researchers have condensed scales that reveal this underutilization of categories (e.g., Engelhard and Chang 2015). Although it could be that students in different contexts that contain less features of authentic writing instruction will use the full scale, we wondered if perhaps there are not as many gradations for perceived authenticity as six. Future research should experiment with using a three-point scale and investigate if this structure improves model-data fit.

As another possible solution some items may be dropped or re-written. For example, Item 2, "I enjoy writing in my ELA class," could be misfitting because it is measuring enjoyment versus authenticity. Potentially, students can rate classroom writing instruction as highly authentic yet not enjoy it. This item may be dropped in future administrations. Also, Item 1, "The writing that I do in my ELA class is related to my life outside of class" may be endorsed by students who are perceiving the authenticity of their classroom writing instruction at very different levels, as long as there is some connection to their lives outside of the classroom. This could be rewritten as "The writing that I do I my ELA class is strongly connected to my life outside of class," which may make this item harder to endorse for those in low to medium authenticity classrooms and thus yield better model-data fit.

Finally, three items exhibiting some misfit may need to be revised: Item 3, "ELA writing assignments relate to topics I care about in the world;" Item 4, "People other than my teacher read the papers I write for school;" and Item 12, "I discuss the topics of my writing assignments with friends." The items identify features that may increase authenticity for some but not for all students. Thinking about the subjective nature of authenticity, certain students may value external readers (Item 4) when considering the overall authenticity of a task, while other students may not value this element of writing instruction. Similarly, discussing topics of writing assignments with friends (Item 12) or writing about issues of global import (Item 3) may matter more for some students than others. For these three items that exhibit misfit, future qualitative research with students investigating factors of authenticity may help refine the language so that these specific factors can represent broader, more universal features of authenticity. Thus, a major recommendation moving forward in addition to the minor modifications suggested here is to pair student and teacher interviews with MPAW Scale use. Because of the complexity of the construct and the lack of research on measuring authenticity, qualitative data can be used to support the future revisions of the MPAW Scale and interpretation of MPAW Scale use.

Conclusion

Based on our analyses, we believe the MPAW Scale is potentially a useful tool for examining overall impressions of authenticity of writing. This study provides evidence that the scale is reliable, as well as some validity evidence for use with students of color in an urban setting. Although the participants in the current study did not represent the full range of ethnicity and socioeconomic status in the United States, one of our major goals in our program of research is to examine perceived authenticity for historically underserved students, and the piloting of the scale with this particular subgroup was a strategic decision. However, future research is needed in different contexts to examine if the scale operates differently (e.g., differential item functioning) for different subgroups of students.

These analyses serve as the base for future studies that will examine teacher and student perspectives on writing instruction and assessment. Because consequential validity is a key facet of a holistic view of validity, an instrument that can easily and accurately capture student perceptions of the authenticity of writing instruction can be a useful source of validity evidence, especially when paired with qualitative data to support interpretation of quantitative data. Future work will include qualitative data sources, such as student and teacher interviews to help interpret authenticity data. If large-scale writing assessments are linked to writing instruction characterized by low authenticity on scales such as the MPAW Scale, these assessments may need to be revised.

Honoring and prioritizing consequences aligns with a vision of writing assessment research that considers students as primary stakeholders in the assessment process (Behizadeh and Engelhard 2014; Guba and Lincoln 1989; Slomp et al. 2014) who should be protected from outcomes (regardless of intention) that are damaging to students' affective or cognitive development, as well as their academic achievement. Researchers have often determined what counts as authentic for students, rather than asking students themselves what they need for authentic education and authentic writing instruction. Students are important stakeholders in large-scale assessments, and their perspectives are underrepresented in discussions of reliability, validity, and fairness of score meaning and use. Soliciting student perspectives on authenticity or other affective variables during the validation process will offer another source of validity evidence that can be used to examine consequential validity, such as the access of historically underserved students to engaging, authentic writing instruction.

Appendix

Comparison of PAW Scale and MPAW Scale

PAW Scale	Modified PAW Scale
1. This writing assignment was relevant and/or meaningful to my life outside of class	1. The writing I do in my English language arts class is related to my life outside of class
2. People other than my teacher will want to read the paper I wrote	4. People other than my English teacher read the papers I write for school
3. Writing this paper was a good learning experience	15. Writing in my English language arts class is making me a better writer
4. I can make connections between this paper and events or issues in the world that I care about	3. English language arts writing assignments relate to topics I care about in the world
6. I will use what I learned writing this paper to write other papers	5. I will use what I am learning about writing to write other papers in the future
7. I have discussed or will discuss the topic of this paper with family members	6. I discuss the topics of my English language arts writing assignments with my family
8. I enjoyed writing this paper	2. I enjoy writing in my English language arts class
9. I think knowing how to write a paper like this one will be important to know in my life	7. What I am learning about writing is important to know in my life
10. Writing this paper was important to me	8. English language arts writing assignments are important to me
11. This paper connects to my personal interests	9. Writing in my English language arts class connects to my personal interests
12. People who read this paper will change their opinions, actions, or feelings	10. People who read my English language arts writing assignments will change their opinions, actions, or feelings
13. I am proud of what I wrote	11. I am proud of what I write in my English language arts class
14. Writing this paper helped me to understand the topic better	16. Writing in my English language arts class helps me to understand topics better
15. I have discussed or will discuss the topic of this paper with friends	12. I discuss the topics of my writing assignments with friends
16. I will use the skills that I learned writing this paper later in my life	13. I am gaining writing skills that I will use later in my life in my English language arts class
17. Writing this paper helped me to develop my thoughts, opinions, or beliefs	14. Writing in my English language arts class helps me develop my thoughts, opinions, or beliefs
5. This paper connected to something I recently saw on TV or the internet*	

*This item is not included in the modified scale

References

Applebee, A. N., & Langer, J. A. (2011). A snapshot of writing instruction in middle schools and high schools. *English Journal, 100*(6), 14–27.

Ashton, S. (2010). Authenticity in adult learning. *International Journal of Lifelong Education, 29* (1), 3–19.

Au, W., & Gourd, K. (2013). Asinine assessment: Why high-stakes testing is bad for everyone, including English teachers. *English Journal, 103*(1), 14–19.

Ball, A. F., & Ellis, P. (2008). Identity and the writing of culturally and linguistically diverse students. In C. Bazerman (Ed.), *Handbook of research on writing: History, society, school, individual, text* (pp. 499–513). New York, NY: Lawrence Erlbaum.

Behizadeh, N. (2014). Adolescent perspectives on authentic writing. *Journal of Language and Literacy Education, 10*(1), 27–44. Retrieved from http://jolle.coe.uga.edu.

Behizadeh, N. (2015). Engaging students through authentic and effective literacy instruction. *Voices from the Middle, 23*(1), 40–50.

Behizadeh, N., & Engelhard, G. (2014). Development and validation of a scale to measure perceived authenticity in writing. *Assessing Writing, 21*, 18–26.

Common Core State Standards Initiative. (2015). *English language arts standards.* Retrieved from http://www.corestandards.org/ELA-Literacy/.

Council of Chief State School Officers. (2011). *InTASC model core standards at a glance.* Retrieved from http://www.ccsso.org/Resources/Publications/InTASC_Standards_At_a_Glance_2011.html.

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future*. New York, NY: Teachers College Press.

Dyson, A. H., & Freedman, S. W. (2003). Writing. In J. Flood, D. Lapp, J. R. Squire, & J. M. Jensen (Eds.), *Handbook of research on teaching the English language arts* (2nd ed., pp. 967–992). Mahwah, NJ: Lawrence Erlbaum.

Engelhard, G. (2013). *Invariant measurement: Using Rasch in the social, behavioral, and health sciences*. New York, NY: Routledge.

Engelhard, G., & Chang, M. (2015). Examining the teachers' sense of efficacy scale at the item level with Rasch measurement model. *Journal of Psychoeducational Assessment*, 1–15.

Fisher, M. T. (2007). *Writing in rhythm: Spoken word poetry in urban classrooms*. New York, NY: Teachers College Press.

Freire, P. (1970/2000). *Pedagogy of the oppressed* (M. B. Ramos, Trans.). New York, NY: Continuum. (Original work published 1970).

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Haertel, E. H., Moss, P. A., Pullin, D. C., & Gee, J. P. (2008). Introduction. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 1–16). Cambridge: Cambridge University Press.

Hillocks, G, Jr. (2011). Commentary on "Research in secondary English, 1912–2011: Historical continuities and discontinuities in the NCTE imprint". *Research in the Teaching of English, 46* (2), 187–192.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *English Journal, 91*(1), 79–87.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review, 64*(1), 76–95.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Morrell, E. (2008). *Critical literacy and urban youth: Pedagogies of access, dissent, and liberation*. New York, NY: Routledge.
National Council of Teachers of English and International Reading Association. (2012). *Standards for the English language arts*. Retrieved from http://www.ncte.org/standards/ncte-ira.
Newmann, F. M., Marks, H. M., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education, 104*(4), 280–312.
Partnership for 21st Century Learning. (n.d.) Framework for 21st century learning. Retrieved from http://www.p21.org/our-work/p21-framework.
Purcell-Gates, V., Anderson, J., Gagne, M., Jang, K., Lenters, K. A., & McTavish, M. (2012). Measuring situated literacy activity: Challenges and promises. *Journal of Literacy Research, 44*(4), 396–425.
Purcell-Gates, V., Duke, N. K., & Martineau, J. A. (2007). Learning to read and write genre-specific text: Roles of authentic experience and explicit teaching. *Reading Research Quarterly, 42*(1), 8–45.
Sisserson, K., Manning, C. K., Knepler, A., & Jolliffe, D. A. (2002). Authentic intellectual achievement in writing. *English Journal, 91*(6), 63–69.
Seunarinesingh, K. (2010). Primary teachers' explorations of authentic texts in Trinidad and Tobago. *Journal of Language and Literacy Education [Online], 6*(1), 40–57.
Slomp, D. H., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. *Research in the Teaching of English, 48*(3), 276–302.
Splitter, L. J. (2009). Authenticity and constructivism in education. *Studies in Philosophy and Education, 28*, 135–151.
Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press. (Original work published 1960).
Watanabe, M. (2007). Displaced teaching and state priorities in a high-stakes accountability context. *Educational Policy, 21*(2), 311–368.
Winn, M. T., & Johnson, L. (2011). *Writing instruction in the culturally relevant classroom*. Urbana, IL: National Council of Teachers of English.
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch measurement transactions, 8*(3), 370.

Author Biographies

Dr. Nadia Behizadeh is an Assistant professor of adolescent literacy at Georgia State University in the United States. Her research explores authentic and culturally sustaining literacy instruction and assessment for adolescent learners, with a focus on increasing the impact of adolescent student writing.

Dr. George Engelhard, Jr. is a professor of educational measurement and policy at The University of Georgia in the United States. He is a fellow of the American Educational Research Association. His latest book is Invariant measurement: Using Rasch models in the social, behavioral, and health sciences (New York: Routledge).

Multifaceted Rasch Analysis of Paired Oral Tasks for Japanese Learners of English

Rie Koizumi, Yo In'nami and Makoto Fukazawa

Introduction

The use of multifaceted Rasch measurement (MFRM) has prevailed in the assessment field, especially in assessing second language (L2) speaking and writing, which involves complex interactions between test takers, tasks, raters, rating scales, and other factors. McNamara and Knoch (2012) describe how Rasch measurement, especially MFRM was adopted in L2 testing communities in the 1990s. Recent applications include Davis (2016) and Aryadousta (2016), both of which investigated the complex nature of L2 speaking assessment.

One type of L2 speaking assessment that has attracted attention from teachers and test developers is a paired oral test (paired oral, hereafter). In paired orals, test takers make pairs and talk with each other rather than an interviewer, and interactions are evaluated by raters. The English ability to interact with various speakers, convey facts precisely, and express one's opinions promptly, while responding to listeners and managing interaction should be fostered and measured, since this ability is one of the determinants of success in today's globalized world. Although there are benefits to using paired orals in L2 speaking assessment, paired oral-related research and applications are limited in Japan. Koizumi et al. (in press) developed and examined a paired oral test consisting of four tasks for Japanese university students learning L2 English. This study builds on Koizumi et al. and expands the number of paired oral tasks calibrated on a logit scale and examines its usefulness.

R. Koizumi (✉)
Juntendo University, Chiba, Japan
e-mail: rkoizumi@juntendo.ac.jp

Y. In'nami
Chuo University, Tokyo, Japan

M. Fukazawa
University of the Ryukyus, Okinawa, Japan

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_6

Previous Studies on Paired Orals

The literature has shown that paired orals have unique values different from other types of speaking assessment with teacher–candidate interaction, especially in two points: First, paired orals can measure oral interaction that is likely observed in natural, real-life conversation when conversation partners have equal status, because the test takers have chances and are required to be active in maintaining the conversation and producing the discourse in cooperation with another test taker of an equal status (e.g., Galaczi and ffrench 2011). Second, paired oral formats are usually similar to pair activities often conducted in communicative classes. Using paired orals can give students a clear message that what they are doing in class is important for good grades and eventually for their future use of English. Similarities between teaching and assessment activities also make it easier for teachers to relate the assessment results for instruction (e.g., Negishi 2015).

Paired orals have two main disadvantages (Negishi 2015). Firstly, test takers' performance and scores can be affected by factors other than their L2 proficiency, such as their own and their partners' L2 proficiency, personality, and degree of familiarity with each other (e.g., Galaczi and ffrench 2011). Second, paired orals tend to have lower reliability across raters and test occasions than examiner–interview and monolog formats, as can be inferred from the study of a group oral test with four test takers discussing topics (Van Moere 2006). Although these are certainly issues that need to be addressed, they may not matter much in low-stakes testing contexts such as classroom assessment, where teachers can assess and judge students' ability on multiple occasions in combination with a few formats.

Paired orals have been examined from the perspective of factors affecting test scores (e.g., Davis 2009; Galaczi 2008, 2014; Kley 2015; Taylor and Wigglesworth 2009, special issue in *Language Testing*) and incorporated into major speaking tests such as the Cambridge English exams (Galaczi and ffrench 2011). However, in Japan, their research and applications are limited, with a few exceptions such as Negishi (2015) and Koizumi et al. (in press). Negishi (2015) compared university students' performances across three formats (i.e., picture description, paired oral, and group oral) using MFRM. She reported that all test formats and raters fit the Rasch model, the paired oral was the most difficult, followed by the picture description, and the group oral, in that order, and test formats and test takers' proficiency levels affected scores. Koizumi et al. (in press) developed a paired oral test and investigated the validity of the interpretation of paired oral test scores, from four viewpoints: First, all tasks and raters fit the Rasch model, with appropriate rating scale properties. Second, high reliability was observed with one task and two raters, or with three tasks and a single rater (with the cutoff score of $\varphi = 0.70$). Third, the test had a unidimensional structure of one factor affecting all scores. Fourth, paired oral scores were moderately correlated with scores of the Test of English as a Foreign Language (TOEFL) Institutional Testing Program (ITP), as predicted in the test development stage.

Current Study

Considering benefits of paired orals, especially in L2 classroom assessment, but limited applications in Japan, it is important to make them accessible to teachers in Japan. For this purpose, we create a bank of tasks whose difficulty levels are measured with a practical rating scale and whose features related to task are set based on Koizumi et al. (in press). To the authors' knowledge, there are no publications describing an attempt to create a large task bank for paired orals in and outside Japan.

The current study aims to assess the L2 oral interactive ability of university students at the novice and intermediate levels. Using MFRM as well as structural equation modeling (SEM) and generalizability theory, we examine the following six questions that are associated with aspects of validity (Messick 1996) and inferences required to make a plausible validity argument (Chapelle et al. 2008). These six research questions (RQs) and validity aspects and inferences are shown below in the parentheses. The current study examines essential aspects of validity in order to provide building blocks of evidence for validity of the interpretation and use based on paired oral test scores.

1. Does the test have a unitary factor structure underlying the paired oral? (structural aspect; Explanation inference)
2. Do all tasks and raters fit the Rasch model? (content and structural aspects; Evaluation and Generalization inferences)
3. Do test tasks have a wide range of difficulty and no wide gaps in difficulty? (content and structural aspects; Evaluation inference)
4. Is the difficulty of student cards equal? (generalizability aspect; Generalization inference).
5. Does the holistic rating scale function properly? (structural aspect; Evaluation inference)
6. How many tasks and raters are minimally needed to obtain sufficient reliability? (generalizability aspect; Generalization inference).

RQ1 examines the test structure of the paired oral; assessing only a single dimension (unidimensionality) is a crucial assumption for Rasch analysis to be met. RQ2 and RQ5 address qualities of tasks, raters, and the holistic rating scale used in this study. RQ3 examines the distribution of tasks to determine whether the tasks are sufficient in number and range for assessing novice- and intermediate-level learners of English. We argue that a task bank should have many tasks with a wide range of difficulty and should not have no-task areas on the Rasch logit scale, since we intend to create a bank of tasks useful for teachers to choose from depending on their test purposes and on their target learners who may have different levels of ability. RQ4 examines comparability of student cards. In the paired oral, one student receives a student card either for Student A or B. We intend to make each card's difficulty level equal and examine this in the RQ. RQ6 inspects the degree of reliability that the paired oral can assure depending on the number of tasks and raters.

Method

Participants

A total of 190 students from three private universities in Japan participated. Their majors were technology, medicine, or English. Their L2 proficiency levels ranged mostly from novice to intermediate. Most of the participants were originally from Japan, so their mother tongue was Japanese. A majority were first-year students who had studied English for at least 6 years at secondary school. Some students were from other countries, but we included them because we intended to create a test for classrooms at Japanese universities that have some overseas students. In their English lessons, the students were instructed to make a pair by themselves to mitigate the influence of familiarity. We did not control or examine the effect of proficiency this time but this should be addressed in the future.

Materials and Procedures

The test included an easy warm-up task and 11 assessment tasks—seven role-plays and four discussions. We used four tasks (Tasks 1–4) similar to the ones used in Koizumi et al. (in press) but modified some instructions by providing specific contexts for the conversation and more familiar place names, and created seven new tasks with familiar topics (see Table 1 for all the tasks). The students were requested to talk for about 2 or 3 min per task. They were not given any planning time and were encouraged to talk in a natural, two-way style with back channeling and eye contact.

After making a pair, students received a student card either for Student A or B, which provided a warm-up and 11 tasks. For raters to identify who was speaking, students were told to begin each task with their name. Tasks were either role-play or discussions. In the role-play task, the card contained a role to play and who should speak first (see Table 1). For example, in Task 9 (Role-play 5: Toothache) Student A needs to begin the conversation, and say that s/he has a terrible toothache; Student B should respond with sympathy and suggest going to see a dentist or take a painkiller; Student A should refuse suggestions at least once and they should continue the conversation. Out of seven role-play tasks, Student A should begin in three, whereas Student B should do so in the remaining four. We intended to make the Student A and B cards comparable in terms of difficulty.

The order of performing tasks was partially counterbalanced: Approximately a third of students performed a warm-up task and tasks 1–10 (not 11) in that order. Another third performed a warm-up task, tasks 4–10, and 1–3 (not 11) and the rest performed a warm-up task, tasks 8–10, and 1–7 (not 11). One class performed only tasks 3, 5, 8, 6, 9, 7, 10, and 11, in that order, because of the limited class time.

Table 1 Warm-up tasks and 11 assessment tasks used

Task	Instruction
Warm-up [A]	Talk about (a) brothers or sisters, (b) pets, (c) boys' (or girls') high school or coeducational school, or (d) favorite food (2 min)
1. Club (RP1)[a] [B]	For A: You are in a cooking club. B is considering whether to join it and wants to ask you questions. Use the information below and answer B's questions kindly. *New Cooking Club! Join us and learn to cook some amazing meals! Every Wednesday after lessons, School Hall, 30 members, £5 a term* For B: Ask questions using the keywords below *When? Where? How many members? Cost?* (2 min)
2. Dinner (RP2) [A]	For A: You want to invite B to come to dinner at your house on Friday evening. If B agrees, talk about details. If B declines, you should negotiate with B about a possible date For B: You have another appointment for Friday evening. Say that you are sorry, that you will not be able to attend and the reason. If A suggests another date, agree if it is okay and talk about details (2 min)
3. Hobby (D1)	Have a conversation related to hobbies (e.g., *sports, clubs, last weekend, Golden Week*) (2 min)
4. Trip (D2)[b]	A and B have agreed to go on a trip together. Decide four things to bring for the trip, while asking each other questions *Place and time: Zoo in Hokkaido in January, Purpose: Seeing cute animals, Weather: Very cold at daytime and night* (2 min)
5. Job (RP3)[c] [A]	For A: B is the owner of the shop where you are considering applying for a part-time job. Ask questions using the keywords below *Name/shop? address? what/sell? telephone number? work every day?* For B: Use the information below and answer A's questions kindly *Happy Feet Store. We need a shop assistant to sell children's shoes. £6 per hour, Saturdays only: 9–5.30 pm, 8 Station Road, Phone 766814* (2 min)
6. Movie (RP4) [B]	For A: You are invited by B. Decline the offer at least once and explain the reason. Agree later if you like the suggested plan For B: You invite A for a movie. It can be viewed at a theater nearby and A will surely like it. If A agrees, talk about details. If A disagrees, convince A to see it on another day (2 min)
7. Friends (D4)	Have a conversation related to friends (e.g., *high school, university, part-time job, meet*) (2 min)
8. Date (D5)	Ken is A and B's friends. He is going on a first date with his girlfriend. He has asked A and B the following question. Discuss how A and B will give Ken advice Question: *Ken invited her for a date. Should he pay for everything? How much should he pay if he does not pay all?* (3 min)
9. Toothache (RP5) [A]	For A: You have a toothache. Say how terribly it aches. Refuse B's advice at least once. Agree later and decide what to do if B suggests a plausible plan For B: Show understanding of A's situation. Recommend that A go to the dentist and/or take a painkiller. Convince A by suggesting concrete plans (2 min)

(continued)

Table 1 (continued)

Task	Instruction
10. Driving (RP6) [B]	For A: Listen to B and tell B that you understand his/her feelings. Cheer B up. Give B advice on what to do next For B: You have failed a driving test. Tell A how sad you are. Decide what to do next based on A's advice (2 min)
11. Victory (RP7) [B]	For A: You won the game. B will celebrate the victory. Talk humbly. Say that you would like to appreciate those who helped you (anybody is okay) For B: Congratulate A, who won the game. Praise A even when A talks humbly (2 min)

Note ( ) = task type No. [ ] = Who should begin the conversation. RP = Role play. D = Discussion. A/B = Students A/B. Underlined = information shared with A and B. The instructions were originally written in Japanese. Students were instructed to continue the related conversation after finishing the assigned task

[a]Derived from Edwards (2008, p. 18–19)

[b]Adapted from Butler and Zeng (2014)

[c]Derived from University of Cambridge ESOL Examinations (2010, pp. 2–4)

Analyses

For rating, we used the same holistic rating scale of 1–3 in Koizumi et al. (in press), which considers interactional effectiveness and linguistic elements such as task achievement, fluency, accuracy, and appropriateness (see Table 2). We created a holistic scale since we weighed practicality over providing detailed feedback to students. We prioritized making a scale that enables teachers to evaluate by listening to the conversation once.

Using the scale, the three authors rated each talk independently. We had a 1-day rater training session assessing five pairs (n = 10), discussing any divergences, and

Table 2 Holistic rating scale

3	*Satisfies adequately* Satisfies the following task point(s). Communicates effectively in English by appropriately participating in turn-taking. Speaks fluently to the extent that the conversation is moving smoothly (Satisfies most of these abovementioned points.) E.g., Task 2 (Role-play 2: Dinner): The person who invites can do so appropriately and continue the related conversation The invited person can say no, apologize, and give reasons for not accepting the invitation appropriately and continue the related conversation
2	*Satisfies to a certain degree* Satisfies some of the task point(s). Communicates adequately in most everyday contexts but can be rather passive in responding and commenting (or mostly speaks alone, dominantly). Due to poor fluency, the conversation does not go smoothly, but the speaker aims to continue the conversation in English
1	*Needs more effort* Satisfies few task point(s). Gives simple responses only when required but is unable to maintain or develop the interaction. Stops the conversation unnaturally and does not make efforts to start it

adding some notes for the scale. We then evaluated the remaining students independently. One of the authors (Rater 1) rated all the remaining 180 students, Rater 2 rated 48, and Rater 3 rated 94. Scores from Raters 2 and 3 were combined and treated as Rating 2, while Rater 1's scores were considered Rating 1.

For MFRM, we used 190 students' scores, 11 tasks, and three raters, with missing values. For SEM and generalizability theory, we used 117 students' scores, 10 tasks, and two ratings, without missing values. The two groups can be considered similar because they had similar means and *SD*s of Rasch ability estimates ($M = 0.43$, $SD = 2.41$, $N = 190$; $M = 0.53$, $SD = 2.38$, $n = 117$). For SEM, we used a robust weighted least squares (WLSMV) estimation method and the software M*plus* (Muthén and Muthén 2014) since the scores were on an ordered scale of 1–3 (RQ1). For MFRM, we used the rating scale model in an MFRM program, Facets (Linacre 2014; RQ2 to RQ5), to estimate the test takers' ability, task difficulty, rater severity, and rating scale. We performed generalizability theory using GENOVA (Center for Advanced Studies in Measurement and Assessment 2013) to calculate the number of tasks and raters needed to obtain highly consistent scores (RQ6).

Results and Discussion

Does the Test Have a Unitary Factor Structure Underlying the Paired Oral?

SEM allows us to construct models hypothesizing relationships between observed and latent variables, based on substantive theory and previous results, and to test whether these models fit the data well (see, e.g., Ockey and Choi 2015; Kline 2010). We hypothesized two models: a unitary model of one factor of oral interactive ability representing 11 tasks (Model 1) and an alternative model of two correlated factors (role-play and discussion abilities) representing two tasks each, as task formats may affect the structure (Model 2). For both models, we used Ratings 1 and 2 for each task (see Fig. 1).

Table 3 shows fit statistics for the unitary (Model 1) and correlated models (Model 2). Although the chi-square statistic was statistically significant ($\chi^2 = 371.921$, $df = 170$, $p < 0.01$; $\chi^2 = 371.612$, $df = 169$, $p < 0.01$) for both models, some indices showed a good fit (CFI = 0.95, TLI = 0.95), while others showed only a moderate fit (RMSEA = 0.10 [0.09, 0.12] and WRMR = 1.21). Model 2 was particularly problematic since its covariance matrix was not positive definite. One reason may be a correlation greater than or equal to 1 between two latent variables. The standardized path between the two-ability factors was 1.002. Model 2 was excluded from further consideration.

Model 1 was revised based on theory and modification indices. A revised model —Model 3—explained the data well (CFI = 0.97, TLI = 0.97, RMSEA = 0.08 [0.06, 0.10], WRMR = 0.99), with the parameter estimates presented in Table 4.

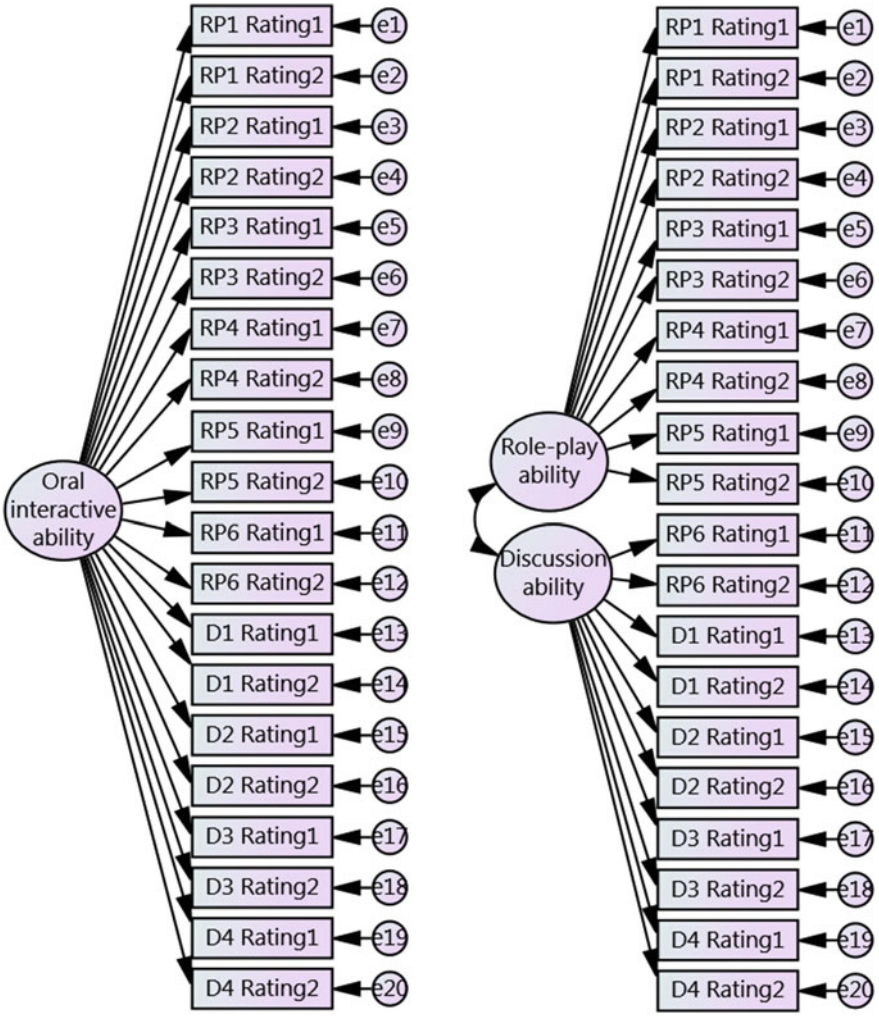


Fig. 1 Model 1 is on the *left* and Model 2 is on the *right*. Each observed variable is labeled by task type and rating. For example, RP1 Rating1 and D1 Rating1 refer to Role Play 1 Rating 1 and Discussion 1 Rating 1, respectively. See Table 1 for RP and D tasks

This suggests that the paired oral is considered to measure a single trait, which we interpret as oral interactive ability, which accords well with the intended test construct. The unitary structure adopted was the same overall as in Koizumi et al. (in press).

We also conducted MFRM and found that 45.05 % of the score variance was explained by Rasch measures, which also suggests unidimensionality of the structure. This percentage of the variance explained by Rasch measures was a little smaller than but similar to Koizumi et al. (in press; 57.90 %).

Table 3 Model fit indices

	χ^2 (*df*) *p*	CFI	TLI	RMSEA [90 %CI]	WRMR
Criteria	$p > 0.05$	>0.90	>0.90	<0.08	<1.00
Model 1: Unitary	371.921 (170) < 0.01	0.95	0.95	0.10 [0.09, 0.12]	1.21
Model 2: Two abilities correlated	371.612 (169) < 0.01	0.95	0.95	0.10 [0.09, 0.12]	1.21
Model 3: Unitary + correlated errors	283.184 (163) < 0.01	0.97	0.97	0.08 [0.06, 0.10]	0.99

Note $N = 117$. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval; WRMR = Weighted Root Mean Square Residual. The criteria were overall based on Byrne (2012) and Yu (2002, as cited in Wang and Wang 2012)

Table 4 Standardized parameter estimates for Model 3

	Path from oral interactive ability	Correlated error	Standard error	R^2
RP1 Rating1	0.75		0.05	0.56
RP1 Rating2	0.87		0.04	0.76
RP2 Rating1	0.76		0.05	0.58
RP2 Rating2	0.86		0.04	0.73
RP3 Rating1	0.54		0.08	0.29
RP3 Rating2	0.76		0.05	0.57
RP4 Rating1	0.73		0.06	0.53
RP4 Rating2	0.76		0.04	0.58
RP5 Rating1	0.74		0.05	0.54
RP5 Rating2	0.93		0.02	0.87
RP6 Rating1	0.75		0.06	0.56
RP6 Rating2	0.87		0.03	0.77
D1 Rating1	0.75		0.06	0.56
D1 Rating2	0.77		0.05	0.60
D2 Rating1	0.72		0.05	0.52
D2 Rating2	0.72		0.05	0.51
D3 Rating1	0.77		0.05	0.60
D3 Rating2	0.86		0.03	0.73
D4 Rating1	0.77		0.05	0.60
D4 Rating2	0.88		0.02	0.78
RP2R1 and RP1R1		0.71		
RP4R1 and RP3R1		0.49		

(continued)

Table 4 (continued)

	Path from oral interactive ability	Correlated error	Standard error	R^2
RP6R1 and RP5R1		0.63		
D2R2 and D2R1		0.55		
D2R2 and RP4R2		0.44		
RP4R2 and RP3R2		0.42		
D1R2 and RP2R2		0.54		

Note Each observed variable is labeled by task type and rating. For example, RP1 Rating1 and D1 Rating1 refer to Role Play 1 Rating 1 and Discussion 1 Rating 1, respectively. The path from RP1 Rating1 is set to 1 for identification. All other factor loadings are statistically significant

Do All Tasks and Raters Fit the Rasch Model?

Figure 2 displays the relationships between ability, task difficulty, rater severity, and rating scale. As seen in Table 5, test takers' abilities spread very widely from −3.67 to 7.51. The task difficulty varied from −1.14 to 1.10, with Task 4 (Discussion 2: Trip) being the most difficult and Task 3 (Discussion 1: Hobby) the easiest. Koizumi et al. (in press) used similar tasks and the order was Task 1 (Role-play 1: Club, the most difficult), 4 (Discussion 2: Trip), 2 (Role-play 2: Dinner), and 3 (Discussion 1: Hobby). Compared to the current study, the order of difficulty of Tasks 4, 2, and 3 was the same. One reason Task 1 had a higher difficulty level than this study was that we added the context (e.g., *You are in a cooking club. B is considering whether to join it.*) of talking about a club; without the context, students must have found it hard to talk in the previous study. Because of the modification, the difficulty seems to have decreased at an appropriate level.

The rater severity differed across raters from −0.66 to 0.60, with Rater 2 as the most severe. Test-taker and task reliability were high (0.91–0.92), which shows consistency of scores across test takers and across tasks. High rater reliability (0.98) indicated that rater severity was different.

The infit mean square statistics between 0.5 and 1.5 were used to judge acceptable model fit (Linacre 2013). However, we did not regard an overfit as problematic (i.e., an infit mean square of below 0.5), because this indicates that the persons, tasks, and raters fit the model too well. We did not also regard an infit mean square between 1.5 and 2.0 as problematic, because it is "unproductive for construction of measurement, but not degrading" (Linacre 2013, p. 270). All the tasks and raters had values within this range, with 0.88–1.22 for the task and 0.83–1.10 for the raters. Furthermore, 15 students (7.89 %, 15/190) had values of less than 0.50 and were considered overfitting students, and 17 (8.95 %, 17/190) had infit mean squares of more than 1.5 and were considered underfitting students, but

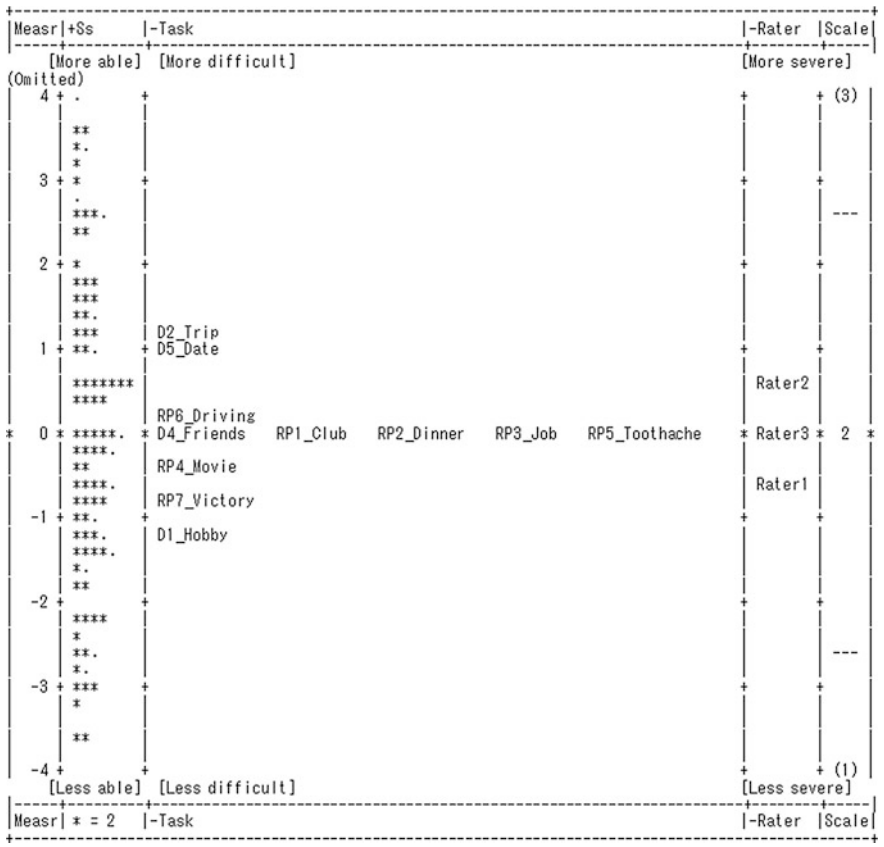


Fig. 2 Wright map for participants (N = 190), tasks (k = 11), raters (n = 3), and the rating scale. Ss = participants; * = 2 participants; . = 1 participant. Fifteen participants with measures of above 4.00 were omitted from the figure. Higher values mean higher ability in the second column, more difficult tasks in the third column, and more severe raters in the fourth column

Table 5 Descriptive statistics for the three facets

Test takers	Logit: M = 0.43, SD = 2.41; Min = −3.67; Max = 7.51 Fair average (on a scale of 1–3): M = 2.07; SD = 0.43; Min = 1.24; Max = 2.99 Reliability = 0.91; Separation = 3.11; Strata = 4.48 Infit mean squares: M = 0.98; SD = 0.41; Min = 0.04; Max = 2.13
Tasks	M = 0.00; SD = 0.62; Min = −1.14; Max = 1.10 Task reliability = 0.92; Separation = 3.34; Strata = 4.79 Infit mean squares: M = 0.99; SD = 0.10; Min = 0.88; Max = 1.22
Raters	M = 0.00; SD = 0.52; Min = −0.66; Max = 0.60 Rater reliability = 0.98; Separation = 6.34; Strata = 8.79 Interrater agreement = 69.6 % (1099/1580); Expected agreement = 64.6 % Infit mean squares: M = 1.00; SD = 0.12; Min = 0.83; Max = 1.10

Note The population *SD*s and reliability for the population with extremes are presented

one test taker had 2.13 and he somehow performed inconsistently across tasks, performing well in one difficult task and worse in an easy task. However, this was only one highly underfitting test taker. The fit of tasks and raters was all appropriate, which was in line with Koizumi et al. (in press).

Do Test Tasks Have a Wide Range of Difficulty and no Wide Gaps in Difficulty?

To see the distribution of tasks available in the test, we examined a task strata statistic, as Linacre (2013) recommends, because we statistically hypothesized that the measure distribution is not normal due to the many items at the peripheral end. We also expected that high and low difficulty levels in task measures are derived because of high and low task difficulty. The task strata was 4.79, which means 11 tasks could be classified into at least four different levels of task difficulty. Although this satisfied a minimum required level, we hope that we can differentiate each level into a few more, so higher task strata would be ideal.

Figure 2 shows that tasks were spread far less widely (range = 2.24, from −1.14 to 1.10) than test takers' abilities (range = 11.18, from −3.67 to 7.51) and that we should have more tasks at higher and lower ends of the scale, that is, more and less difficult tasks. Figure 2 also demonstrates the existence of some gaps on the logit scale. However, as seen in Table 6, most gaps were within the standard error of measurement and were regarded as not very substantive. For example, Tasks 11 (Role-play 7: Victory) and 3 (Discussion 1: Hobby) have a task difficulty of −0.77 and −1.14 respectively, but 68 % confidence intervals (CIs) overlapped (−1.18 to −0.36 and −1.28 to −1.00). There were two cases with different values beyond the standard error: between Tasks 8 (Discussion 5: Date) and 10 (Role-play 6: Driving), and between Tasks 2 (Role-play 2: Dinner) and 6 (Role-play 4: Movie). We can also argue that when we used 95 % CI, there was only a gap in the former case (0.71 to 1.21 and −0.01 to 0.49, not shown in Table 6 but calculated using Measure ± 1.96 * SE), whereas there was an overlap in the latter (−0.34 to 0.20 and −0.63 to −0.13). Nevertheless, we decided to use 68 % CI to strictly improve our test. These two gaps in between as well as at the higher and lower ends can be modified in a future revision by adding tasks with such difficulty levels.

Is the Difficulty of Student Cards Equal?

As explained in the Method section, in the test, a student received a student card for Student A or B, and Students A and B made a pair. We compared the students' ability estimates across the two groups (Students A vs. B groups) but found no significant difference between the groups with the effect size being negligible

Table 6 Task measurement report

Task	Total count	Observed average	Fair average	Measure (logit)	Model SE	Infit MnSq	68 %CI (Measure ± SE)
D2: Trip	302	1.91	1.84	1.1	0.13	1.13	0.97 to 1.23
D5: Date	342	1.97	1.87	0.96	0.13	1.06	0.83 to 1.09
RP6: Driving	310	2.12	1.98	0.24	0.13	1.00	0.11 to 0.37
D4: Friends	339	2.12	2.00	0.10	0.13	0.95	−0.03 to 0.23
RP1: Club	282	2.13	2.01	0.03	0.14	0.93	−0.11 to 0.17
RP9: Toothache	335	2.15	2.01	−0.02	0.13	0.97	−0.15 to 0.11
RP3: Job	345	2.16	2.02	−0.05	0.13	1.22	−0.18 to 0.08
RP2: Dinner	280	2.15	2.02	−0.07	0.14	0.89	−0.21 to 0.07
RP4: Movie	346	2.21	2.07	−0.38	0.13	0.88	−0.51 to −0.25
RP11: Victory	39	2.62	2.13	−0.77	0.41	0.96	−1.18 to −0.36
D1: Hobby	314	2.36	2.19	−1.14	0.14	0.92	−1.28 to −1.00
Mean	294	2.17	2.01	0.00	0.16	0.99	–
SD[a]	83.90	0.18	0.01	0.62	0.08	0.10	–

Note SE = Standard error. MnSq = Mean squares. CI = Confidence interval
[a]Population

(Student A: $M = 0.30$, $SD = 2.43$, $n = 95$; Student B: $M = 0.57$, $SD = 2.41$, $n = 95$; $t = -0.78$, $df = 187.99$, $p = 0.44$, $d = -0.11$, 95 % CI = −0.40 to 0.17). Thus, we can conclude that the difficulty level of student cards is considered equal. It should be noted that this result came from a situation where tasks requiring Student A to speak first are used almost the same number of times as tasks requiring Student B to do so; when teachers select tasks from the task pool, they may need to consider the balance of tasks from this perspective.

Does the Holistic Rating Scale Function Properly?

We analyzed functions of the rating scale based on Bond and Fox (2007). Table 7 indicated that results of the scale almost satisfied the criteria: There were more than 10 ratings at each level (420–1960). Thresholds, or difficulty estimates for choosing one level over another (e.g., −1.99 from levels 1–2) increased as the level increased, and the values of distances between thresholds between neighboring levels were 2.28 and 5.02; the former was between 1.4 and 5.0 logits but the latter was marginally beyond 5.0; we considered this to be minor. The probability curve (Fig. 3)

Table 7 Category statistics for the rating scale

Level	Number of observations (%)	Average measure for test takers at the level	Rasch-Andrich threshold measure (distance), standard error	Outfit mean squares
1	420 (14 %)	−1.99		1.0
2	1960 (65 %)	0.16	−2.51 (2.28), 0.06	1.1
3	645 (21 %)	2.62	2.51 (5.02), 0.06	1.0

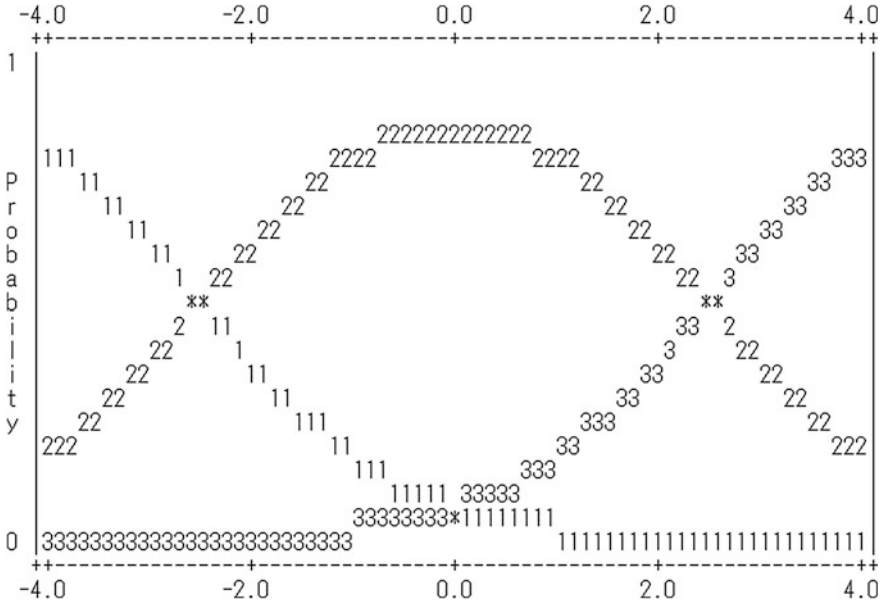


Fig. 3 The probability curve of the scale

had a clear top for Level 2. The level fit statistics were 1.0−1.1, less than 2.0. These results accorded well with the predicted patterns from the Rasch measurement and with Koizumi et al. (in press).

How Many Tasks and Raters Are Minimally Needed to Obtain Sufficient Reliability?

Using generalizability theory (Brennan 2001), we decomposed the test score variance into variance components affected by seven sources: variations of (a) persons' ability (the objects of measurement), (b) task difficulty, (c) raters' severity, (d) person-by-task interaction, (e) person-by-rater interaction, (f) task-by-rater interaction,

Table 8 Estimated variance component and percentage of variance explained

	df	Variance component	Percentage (%)	Standard error
Persons (*p*)	116	0.13	35.81	0.02
Tasks (*t*)	9	0.01	2.45	0.00
Raters (*r*)	1	0.02	4.34	0.01
$p \times t$	1044	0.04	11.71	0.01
$p \times r$	116	0.04	11.31	0.01
$t \times r$	9	0.00	0.71	0.00
Residuals ($p \times t \times r$, *e*)	1044	0.12	33.67	0.01

and (g) the residual, consisting of the person-by-task-by-rater interaction and random errors, in the generalizability (G) study. We considered tasks and raters (which are actually ratings, with Rating 1 [scores from Rater 1] and Rating 2 [scores from Raters 2 and 3 combined]) as random facets. This method is often used for data in which not all raters evaluate all task responses (Lin 2014).

Table 8 shows the percentages of variance explained by the seven sources. The results suggest that the largest variability was explained by the persons (35.81 %), followed by the residual (33.67 %), and, to a lesser degree, by person-by-task (11.71 %) and person-by-rater (11.31 %) interactions. The percentages explained by tasks, raters, and task-by-rater interaction were marginal, ranging from 0.71 to 4.34 %. This suggests that the tasks and raters had similar levels of difficulty and severity. This appears in contrast to results from MFRM stating that task difficulty differed across tasks. However, MFRM results do not show the impact of tasks and raters on scores, and G study results showed that the impact was limited. The pattern in G study was almost the same as in Koizumi et al. (in press) except that the percentage of person-by-task interaction (11.71 %) was larger in the current study than in Koizumi et al. (5.79 %), probably because of an increased number of tasks.

Using the decision (D) study, we investigated how test reliabilities change depending on the number of tasks and raters. We used phi coefficients (Φ), which are used for an absolute decision, but results of generalizability (G) coefficients, for a relative decision, were also presented for interested readers. We employed a criterion of $\Phi = 0.70$ or more, considering the use in low-stakes classroom assessment. Table 9 showed that when one rater evaluates the test, even the use of ten tasks does

Table 9 Phi coefficient (Φ) and generalizability coefficient (in the parenthesis) in decision studies (p × t × r design)

	1 task	2 tasks	3 tasks	4 tasks	5 tasks	6 tasks	7 tasks	8 tasks	9 tasks	10 tasks
1 rater	0.36 (0.39)	0.47 (0.51)	0.53 (0.58)	0.56 (0.61)	0.59 (0.64)	0.60 (0.65)	0.61 (0.67)	0.62 (0.68)	0.63 (0.69)	0.64 (0.69)
2 raters	0.48 (0.51)	0.60 (0.64)	0.66 (0.70)	0.70 (0.74)	0.72 (0.76)	0.73 (0.77)	0.74 (0.79)	0.75 (0.80)	0.76 (0.80)	0.77 (0.81)
3 raters	0.54 (0.57)	0.67 (0.70)	0.72 (0.76)	0.75 (0.79)	0.78 (0.81)	0.79 (0.83)	0.80 (0.84)	0.81 (0.84)	0.82 (0.85)	0.82 (0.86)

Note Underlined = 0.70 or above

not lead to high reliability; when two raters join, at least four tasks are needed to obtain reliable scores; when three raters evaluate, at least three tasks are needed. In classroom assessment, usually one rater is available and in this case, a teacher may need to know that paired orals tend to have low reliability and to use as many tasks as possible. When two raters are available, the required number of tasks is reduced to four and this may be manageable. Koizumi et al. (in press) showed that conditions of one task with two raters, and three tasks with a single rater would produce sufficient reliability. This seems to indicate that when we increase tasks in the task bank, we should check the number of tasks and raters needed because this increase may change the impact of related factors on test scores.

Conclusion

We investigated six aspects related to the validity of the interpretation based on paired oral scores. We found that the structure of our paired oral has a unitary dimension, all tasks and raters fit the Rasch model, test tasks had a moderately wide difficulty range with gaps in between and at the higher and lower ends, the difficulty of student cards was equal, the holistic rating scale functioned properly, and the number of tasks and raters minimally needed to obtain sufficient reliability was at least four tasks with two raters and three tasks with three raters.

The results we obtained in this study were generally positive and as expected in the test developing stage. Major unexpected parts were the existence of gaps in between and at higher and lower ends of the scale, and they will be addressed and rectified in future research. We will also transcribe actual conversations and qualitatively examine relationships between linguistic functions intended to be elicited and those actually observed in the conversation. This information will help us identify what type of tasks should be included in the task bank together with the construct intended and the difficulty information that we obtained in the current study.

Our results will provide teachers with crucial information on how to use paired orals in their classroom. Moreover, we mainly used multifaceted Rasch measurement (MFRM), along with some auxiliary methods (structural equation modeling and generalizability theory) for the validation of our paired oral. MFRM has helped us identify strengths and weaknesses of our test and suggested improvements. The methods we used would be useful for other contexts where test takers, tasks, and raters are involved.

Acknowledgement This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C), Grant Number 26370737.

References

Aryadousta, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly, 13*, 1–24. doi:10.1080/15434303.2015.1133626.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessment. *Language Assessment Quarterly, 11*, 45–75. doi:10.1080/15434303.2013.869814.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.

Cambridge ESOL Examinations (2010). *Speaking test preparation pack for Key English Test*. Cambridge, U.K.: Author.

Center for Advanced Studies in Measurement and Assessment (University of Iowa, College of Education). (2013). *GENOVA suite programs*. Retrieved from http://www.education.uiowa.edu/centers/casma/computer-programs#8f748e48-f88c-6551-b2b8-ff00000648cd.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language.™*. New York, NY: Routledge.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*, 367–396. doi:10.1177/0265532209104667.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*, 117–135. doi:10.1177/0265532215582282.

Edwards, L. (2008). *Common European Framework assessment tests.* London, U.K.: Mary Glasgow Magazines (Scholastic).

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*, 89–119. doi:10.1080/15434300801934702.

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*, 553–574. doi:10.1093/applin/amt017.

Galaczi, E., & ffrench, A. (2011). Context validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 112–170). Cambridge, UK: Cambridge University Press.

Kley, K. (2015). *Interactional competence in paired speaking tests: Role of paired task and test-taker speaking ability in co-constructed discourse.* Unpublished Ph.D. dissertation, University of Iowa, U.S. Retrieved from http://ir.uiowa.edu/etd/1663/.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Koizumi, R., In'nami, Y., & Fukazawa, M. (in press). Development of a paired oral test for Japanese university students. *British Council New Directions in Language Assessment: JASELE Journal Special Edition.*

Lin, C.-K. (2014). *Treating either ratings or raters as a random facet in a performance-based language assessments: Does it matter?* CaMLA Working Papers 2014-01. Cambridge Michigan Language Assessments. Retrieved from http://www.cambridgemichigan.org/sites/default/files/resources/workingpapers/CWP-2014-01.pdf.

Linacre, J. M. (2013). *A user's guide to FACETS: Rasch-model computer programs (Program manual 3.71.0).* Retrieved from http://www.winsteps.com/a/facets-manual.pdf.

Linacre, J. M. (2014). Facets: Rasch-measurement computer program (Version 3.71.4) [Computer software]. Chicago: MESA Press.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*, 555–576. doi:10.1177/0265532211430367.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256. doi:10.1177/026553229601300302.

Muthén, L., & Muthén, B. (2014). M*plus* (Version 7.2) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Negishi, J. (2015). Effects of test types and interlocutors' proficiency on oral performance assessment. *Annual Review of English Language Education in Japan, 26*, 333–348.

Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly, 12*, 305–319. doi:10.1080/15434303.2015.1050101.

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26*, 325–339. doi:10.1177/0265532209104665.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*, 411–440. doi:10.1191/0265532206lt336oa.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. West Sussex, UK: Wiley.

The Scale of Reflective Process in Social Work Practicum

Hui-Fang Chen and Gloria Hongyee Chan

To ensure that social workers are well equipped for dealing with the complexities and challenges of real-life practice (Dolan et al. 2006; Yip 2006; Ruch 2007), reflection should be an important component of social work (Oltedal 2010). The Social Work Reform Board (2010) in the United Kingdom has developed the Professional Capabilities Framework (PCF) to guide curriculum design for continuous professional development (University of Bedfordshire 2014). The PCF includes "critical reflection and analysis" (i.e., application of critical reflection and analysis to inform and provide a rationale for professional decision making) as one of the domains of capabilities. Previous studies have also noted that the reflective process is critical for learning about social work and in professional practice as it alleviates negative emotions, results in the development of new perspectives and solutions, and improves professional suitability (Schön 1993; Yip 2006). However, to our best knowledge, there is no quantitative measurement tool for assessing students' reflective processes in fieldwork practicum, especially self-evaluation. This study aimed to develop a self-report scale for assessing students' reflective ability during social work practicums and evaluate its psychometric properties using the Rasch measurement model.

H.-F. Chen (✉)
Department of Applied Social Sciences, City University of Hong Kong, Tat Chee Ave., Kowloon, Hong Kong
e-mail: hfchen@cityu.edu.hk

G.H. Chan
School of Social Sciences, Caritas Institute of Higher Education, Tseung Kwan O, Hong Kong
e-mail: chanhongyee2004@yahoo.com.hk

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_7

The Concept of Reflection

Reflection can be defined as "a process of reviewing an experience of practice in order to describe, analyze, evaluate and so inform learning about practice" (Reid 1993, p. 305). It refers to "the ground or basis for a belief that is deliberately sought and its adequacy to support the belief examined" (Dewey 1910, pp. 1–2). It is a special type of problem solving in which the resulting ideas are linked to each other and lead to an improved outcome (Hatton and Smith 1995). While scholars have used a number of different terms to refer to the concept of reflection, such as reflective thought (Dewey 1910), reflection-in-action (Schön 1983), and reflexivity (D'Cruz et al. 2007), they generally agree that reflection is a cognitive process. Calderhead (1989) further pointed out that reflection will evoke actions to justify and consider consequences, rather than appetitive, blind or impulsive actions.

Reflection can be extremely beneficial for learners, especially social work students, because "cultivating reflexivity can help students to understand themselves better and be prepared for a future career as a social worker" (Chow et al. 2011). In addition to enriching their learning, it can also improve students' teamwork and ensure that they are adequately prepared for the working world. Then, students develop their self-monitoring and professional self-constructive ability, which lead to improvement of their professional actions and knowledge throughout their life (Calderhead 1989).

Reflection involves thinking about an idea and possible consequences that enable further learning. In this way, reflection can be seen as a step-by-step process in which each reflective thought can lead to another, helping students to keep growing and learning. The ultimate purpose of reflection is to find "further facts which serve to corroborate or to nullify the […] belief" (Dewey 1910, p. 9) and to "assist the learner to unearth and unsettle assumptions […] and thus to help identify a new theoretical basis from which to improve and change a practice situation" (Fook 1991, p. 446). In other words, learners, will be able to integrate new knowledge with their experiences to make a better choice and an action, enhancing the overall effectiveness (Rogers 2001).

Three major benefits of reflection have been identified in the literature (Fook and Gardner 2007). This first is the ability to make more informed choices as a result of developing the ability to look at things from a different perspective, analyze them, and determine what could be done differently. The second benefit is that learners place higher values on knowledge they have acquired through personal experiences and reflection as well as on new theories they developed. The third benefit is more creative and effective practice, as students are less reluctant to challenge given theories and interpret them differently, ultimately leading to increased professionalism and improved practice.

The Reflective Process

As mentioned above, reflection is regarded as a process (Reid 1993). Scholars have identified different stages or phases of this process (e.g., Dewey 1933; Mezirow 1991; Rogers 2000). Rogers (2000) identified three steps of the reflective process: (1) identifying a problem and deliberately deciding to find a solution; (2) collecting additional information about the problem; and (3) taking action based on the reflective process. Teachers can act as facilitators during these stages and encourage learners to give accurate descriptions and remain objective during the first phase. In the second phase, teachers can draw learners' attention to the feelings they experienced and help them become more conscious of their thoughts and emotions. During the final phase, they mainly act as supporters and can ask questions to help students obtain deeper insights (Boud et al. 1985).

The reflective process is essential in various aspects of social work. Nathan (1993) and Hughes and Pengelly (1997) have pointed out that the reflective process involves a social worker's reflection upon his/her psychodynamic relationship with the client (e.g., Sheppard 2000; Taylor and White 2001). Social workers have to be aware of personal thoughts, feelings, and emotions that can help them gain insight into a client's experience and determine a response to the situation (Ruch 2000). This means that during the reflective process, the social worker's feelings, personal experience, and cognition will be utilized to resolve current difficulties, avoid uncomfortable feelings, evaluate present and past performance, and search for new perspectives and new solutions (Yip 2006).

The reflective process also enhances a social worker's self-awareness and self-understanding. Schön (1993) has pointed out that how social workers self-evaluate their practice is associated with their professional identity. Hence, the reflective process is seen as related to social workers' psychological state, sense of empathy, intercultural sensitivity, professional suitability, and practice. In addition, Gursansky et al. (2010) mentioned that reflection is an important skill that needs to be specifically taught in social work education through methods such as journaling. The claim is that if social work students are taught how to engage in reflection, it is possible that they will be better prepared for social work practicums. Guidance and support from practicum supervisors can facilitate students' engagement in reflection (Davys and Beddoe 2009).

Approaches to Assessing the Reflective Process

Qualitative methods such as keeping a diary, journal, or learning portfolio are the most common methods for evaluating reflection. A diary is mainly used for self-evaluation and is usually private. This means that it will not be assessed by teachers or supervisors; it is primarily intended to help students keep track of their

own development. A learning portfolio is also a personal document in which learners record their activities and learning outcomes (Brockbank and McGill 1998).

There have been a number of quantitative studies that have used scales to measure reflective ability. Examples of such scales include vignettes (Finch 1987; Regehr et al. 2007), an objective structured clinical examination (OSCE) adapted for social work (Bogo et al. 2013), the Self-reflection and Insight Scale (SRIS) (Grant et al. 2002), and the Reflection Questionnaire (Kember et al. 2000). Although there are tools for measuring reflection, they tend to measure individuals' inclination to engage in reflection (e.g., SRIS and the Reflection Questionnaire) rather than the reflective process. Even though the OSCE and vignettes can shed light on the reflective process in a social work context, they are qualitative assessments, not quantitative scales. Hence, this study aims to develop and validate a scale that can measure the reflective process of social work students in practicum-based learning.

Method

Participants

Students enrolled in a university in Hong Kong and completing two fieldwork practicums in 2013–2014 academic year were invited to participate in the present study. A total of 280 students answered and returned the questionnaires during their first and second fieldwork practicums, and their responses were analyzed in the present study. Students participated in the first practicum during the second college year and the second practicum in the third year. Both the two practicums lasted 17 weeks.

The Reflective Process in Practicum Scale (RPPS)

The Reflective Process in Practicum Scale (RPPS) was developed by a social work practicum team at a university in Hong Kong. The RPPS includes eight items for assessing the frequency students engaged in the reflective process during fieldwork practicums. All items were rated on a 4-point scale where 1 is never, 2 is sometimes, 3 is most of the time, and 4 is always. Table 1 lists the item information.

Six experts were invited to review the scale using a 5-point scale (1 = totally disagree and 5 = totally agree). A high rating on an item indicated strong agreement with the use of the item to assess students' reflective process during practicums. Five experts on the reviewing panel were university faculty members, three of whom were from the social work department and two of whom were from the

Table 1 Item information of the reflection process in practicum scale (RPPS)

		The eight item RPPS					The seven item PPPS				
			Infit		Outfit			Infit		Outfit	
Item ID	Item content	Average item difficulty	MNSQ	T statistics	MNSQ	T statistics	Average item difficulty	MNSQ	T statistics	MNSQ	T statistics
A	Doing and reflecting on my action at the same time	0.23	1.05	1.00	1.05	0.90	0.41	1.16	2.7	1.15	2.6
B	Doing and reflecting on my action afterwards	−0.93	0.86	−2.7	0.92	−1.4	−0.87	0.91	−1.7	0.98	0.30
C	Using knowledge to plan my action and evaluating the result afterwards	0.11	0.85	−2.80	0.85	−2.70	0.28	0.95	−0.90	0.94	−1.0
D	Seeking guidance from my field instructor to plan my action	−0.46	0.90	−1.9	0.88	−2.10	−0.35	0.91	−1.70	0.88	−2.10
E	Seeking corrective feedback from my field instructor on what I have done	−0.43	0.90	−1.8	0.90	−1.80	−0.32	0.94	−1.10	0.93	−1.2
F	Rehearsing what I am going to do with my field instructor and getting coaching from him/her	0.74	1.03	0.60	1.02	0.40	0.98	1.15	2.50	1.15	2.40
G	Implementing changes suggested by my field instructor and evaluating the result	−0.26	0.90	−1.70	0.92	−1.50	−0.13	0.97	−0.60	0.97	−0.40
H	Getting feedback from my clients	1.01	1.47	7.20	1.50	7.30					

Note RPPS The Reflection Process in Practicum Scale; *MNSQ* Mean squares of variance

psychology department. The other expert was a postdoctoral fellow in the social work department. The three social work professors supervised social work students in practicums for a decade and published many related articles. The other two professors and the postdoctoral fellow were involved in test development and validation. The experts were asked to write down their opinions of the items. Their opinions and ratings for each item were used to establish the content validity of the RPPS.

Data Analysis

The rating scale model (Andrich 1978) was implemented in Winsteps 3.75 (Linacre 2012). The following issues were carefully examined, including: (1) the fundamental assumption of unidimensionality, (2) rating category diagnostics on a 4-point Likert scale, (3) the fit of individual items, (4) the range of item difficulty according to four response categories for determining whether the items target students' reflective process during social work practicums, and (5) the potential bias of items related to assessment time.

Because the eight items were developed to measure a unidimensional concept of the reflective process, the underlying construct was therefore assumed to be a single dimension. Principal component analysis (PCA) of the residuals and item fit were examined to against the assumption of unidimensionality. It was assumed that the Rasch dimension explained more than 50 % of variances and the eigenvalue of the first residual factor (the largest component of the unexplained variance in the data) was less than 3, explaining 7 % of variances or less. Patterns in the residuals might indicate that unidimensionality is violated and can be used to determine the pattern of the multidimensional structure of the eight items. Item fit indices were used to examine not only the assumption of unidimensionality but also the appropriateness of individual items for assessing students' reflective process during social work practicums. All items were assumed to have infit and outfit mean squares (MNSQ) within an acceptable range. If the infit and outfit MNSQ of an item falls outside the range of 0.6–1.4 and their corresponding values of t statistics fall outside the range of -2 to $+2$, the item was considered problematic and was carefully checked (Wright et al. 1994). An item with an MNSQ higher than 2.0 degraded the whole measurement and was removed from the scale (Wright et al. 1994).

Rating scale diagnosis was implemented to evaluate the appropriateness of a 4-point scale for assessing students' reflective process. A response category was considered problematic if its outfit MNSQ exceeded 2 with a corresponding t value greater than 2. The Andrich thresholds and category measures were expected to increase monotonically, which means that the threshold and category measures of a lower response category were smaller than the ones of the next category. This means that the average student's reflective ability should increase from a category representing low ability to one representing high ability. If the response categories failed to meet the above requirements, the categories were reorganized.

Because the Rasch model calibrates participants' ability and item difficulty simultaneously, the locations of these elements could be plotted to indicate their relationship. An ideal scale for targeting participants' reflective ability should have very easy to very difficult items to cover the entire range of the reflective process during social work practicums. An assessment tool using a Likert scale usually can cover a wide range of item difficulty and person ability levels. Thus, it was expected that the RPPS with a 4-point scale would target different levels of the reflective ability. We also evaluated test reliability using person (separation) reliability, with 0.7 or above indicating a satisfactory level of reliability.

Differential item functioning (DIF) was implemented to examine item biases related to assessment time. DIF refers to the different probabilities of endorsing an item even though participants have the same reflective ability due to the impact of other factors. In the present study, the questionnaires were administered at two different time points, and it was assumed that participants with the same reflective ability would respond similarly regardless of the time at which they answered the questionnaires. Once DIF occurred, responses to the RPPS were influenced by factors other than reflective ability and changes between the two administrative time points could not be interpreted as improvements or decrements. Noticeable DIF items had a magnitude of 0.5 logits or larger at a nominally significant level of 0.05.

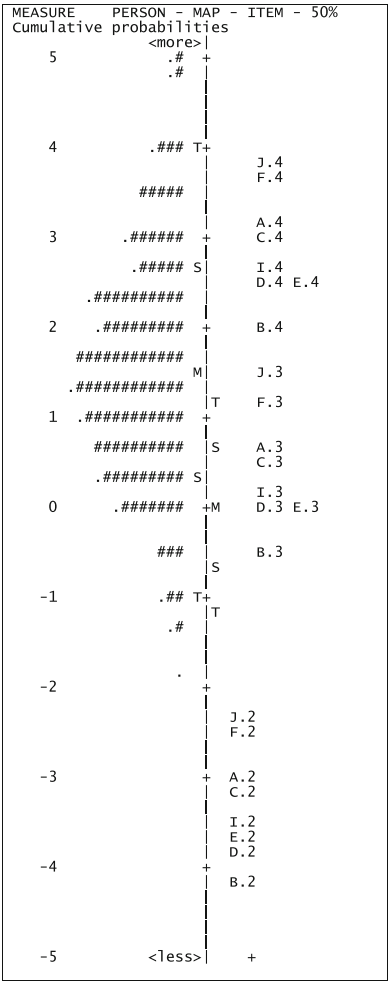
Results

Majority of the participants were female, contributing to 77.2 % of the participants. Their ages ranged between 18 and 22. The average ratings of each item ranged between 3.8 and 4.67, indicating that all experts agreed that the eight items could be used to measure students' reflective process in social work practicums. None of the experts suggested revisions. Thus, combined with the ratings and experts' comments, the content validity of the eight item RPPS was satisfactory.

PCA of residuals indicated that 40 % of variances could be explained by measures, and the eigenvalue of the first contrast was 1.6, explaining 12 % of unexplained variances. The pattern of residuals indicated that there was no second dimension among residuals. Item fit indices suggested that the item "getting feedback from my clients" yielded infit and outfit MNSQ higher than 1.3 at a significant level but had not achieved the removal cut-point of 2 (Table 1).

Rating scale diagnosis showed that the four response categories functioned well for participants to rate their reflective process during practicum. The person–item map indicated that the average participant's ability was 1.54 logits, higher than the average item difficulty (0.0 logits). Participants' ability ranged between −1.91 logits and +6.21 logits, and item difficulty ranged between −1 logit and +1 logit. The items were easier for this group of participants. Using a 4-point scale, the RPPS could target participants with abilities ranging from −5 logits to +5 logits (see Fig. 1a for details). Reliability was 0.7 at an acceptable level. There were no DIF items in the scale, meaning that administrative time did not bias participants' responses.

(a)



(b)

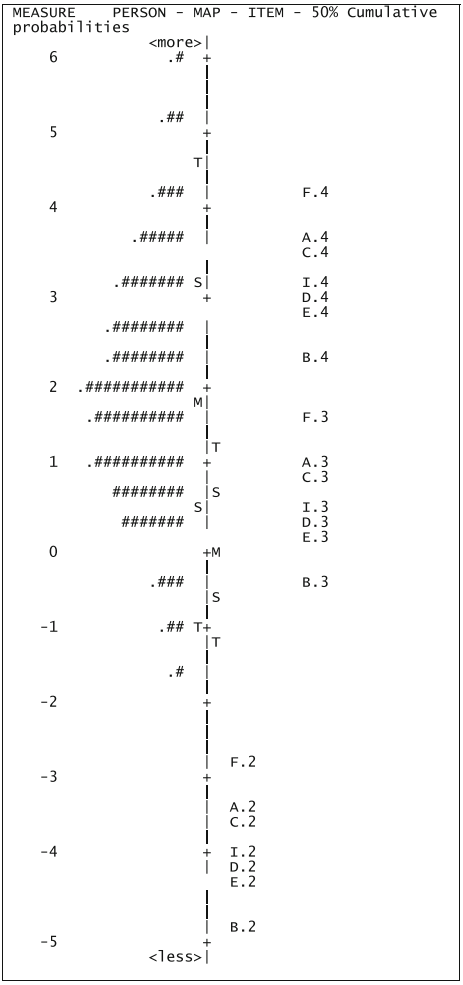


Fig. 1 **a** The eight item scale. *Note* Each "#" indicates 5 people; and each "." indicates 1–4 people. **b** The seven item scale. *Note* Each "#" indicates 6 people; and each "." indicates 1–5 people

Because an item yielded a problematic item fit, the following data analyses compared the impact of removal of this item to determine whether there was any significant improvement before a conclusive suggestion was made. For the seven item RPPS, 41.5 % of variances could be explained by the Rasch dimension, and the eigenvalue of the first contrast was 1.6, explaining 13.6 % of unexplained variances. All seven items showed adequate infit and outfit MNSQ (see Table 1 for details). There were no significant changes in rating scale diagnosis, the person–item relationship (Fig. 1b), reliability, or DIF. There were no significant changes when the item was removed from the RPPS.

Discussion

The practice of reflection is a critical element of the process of learning and acquiring knowledge for social workers because it enables the learner to engage actively with praxis (theory in practice) (Ixer 1999). Social work practicums can be recognized as "a setting designed for the task of learning a practice" (Schön 1987, p. 37) to help students learn by doing and achieve professional requirements. Thus, we developed the RPPS scale for students to evaluate their own practice and performance during a practicum. This approach is an alternative to traditional methods of evaluating a curriculum and designing a training program for social workers.

Students' self-evaluation of their reflective process can be done in an oral or written form, such as in interviews or written journals, but written tasks are more common. They help students to express their thoughts and think about their actions, thus enabling them to reflect on these actions (Hatton and Smith 1995). However, these types of assessments usually take more time to complete, and it might not be easy to use these writings for interpersonal comparisons. The present study implemented a self-evaluation form that takes less than 10 min to complete, shortening the administrative time and providing an overview of reflective ability from students' perspective.

The eight items of RPPS demonstrated satisfactory validity and reliability based on experts' reviews and findings from the Rasch analysis. A panel of six experts strongly agreed with the use of these questions to assess students' understanding of their reflective process during practicums, with ratings between 3.8 and 4.67 on a 5-point scale. All but one item functioned well to measure a single underlying construct with evidence of residual patterns and item fit. Using a 4-point scale with these items can differentiate participants from a very low to a high level of reflective practice during social work practicums. Responses to the eight items were not biased by administrative time, which suggested that further comparisons between different time points should be included in further investigations. The reliability estimate was 0.7, which is an acceptable level. It is conclusive that the eight items could be used to measure students' reflective process.

The item "getting feedback from my clients" showed problematic item fit, so we conducted further investigations to determine the impact of removal of this item. All the remaining seven items showed adequate item fit and reflected a single underlying construct. There were no item biases regarding administrative time, and removal of the item did not harm the reliability estimate. However, this item assesses students' ability to look for and collect information from another perspective (Rogers 2001), which is one of the four essential steps of the reflective process. Thus, we decided not to remove the item from the scale before further investigations are done (such as using a larger sample size to estimate the item fit).

Reflection can happen at both the personal (individual reflection) and interpersonal (corporate reflection) levels. It means that reflective ability can be assessed from self-evaluation and supervisors' evaluations (Smith 2011). Brockbank and

McGill (1998) describe the two forms of the reflective process in a similar way: (1) self-report, in which the learner reflects on his or her own practice, such as in the form of a dialogue or written report, and (2) other-report, in which the learner's reflective activity is reviewed by others in written or oral form to help him or her improve. The present study mainly focused on the personal level and developed a self-evaluation form for students' use. In the future, researchers can develop a questionnaire-type assessment tool for supervisors' use and compare the results of self- and other-reports to see if they are consistent across different perspectives.

Brockbank and McGill (1998) argued that critical learning requires teachers and students to work together. The teacher acts as the facilitator and is responsible for working together with students to enable them to reflect on and better understand what they are learning. Larrivee (2000) further argued that the expectations in classrooms, especially in higher education, are extremely high; thus, teachers need to act as facilitators of learning or reflective practitioners.

Reflective practice is beneficial for teaching and learning, as both students and teachers are able to learn from experiences and improve their practice. By regularly engaging in reflective practice, teachers will understand the process better and thus will be better able to teach it to others (Brockbank and McGill 1998). We believe that the development of scales for assessing reflective ability is essential for higher education, and that students and educators from different disciplines will benefit from the findings.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Bogo, M., Katz, E., Regehr, C., Logie, C., Mylopoulos, M., & Tufford, L. (2013). Toward understanding meta-competence: An analysis of students' reflection on their simulated interviews. *Social Work Education: The International Journal, 32*(2), 259–273.

Boud, D., Keogh, R., & Walker, D. (1985). *Reflection: Turning experience into learning*. London: Kogan Page.

Brockbank, A., & McGill (1998). *Facilitating reflective learning in higher education*. Buckingham: Society for Research into Higher Education/Open University Press.

Calderhead, J. (1989). Reflective teaching and teacher education. *Teaching and Teacher Education, 5*(1), 43–51.

Chow, A. Y. M., Lam, D. O. B., Leung, G. S., Wong, D. F. K., & Chan, B. F. P. (2011). Promoting reflexivity among social work students: The development and evaluation of a program. *Social Work Education, 30*(2), 141–156.

D'Cruz, H., Gillingham, P., & Melendez, S. (2007). Reflexivity, its meaning and relevance for social work: A critical review of the literature. *British Journal of Social Work, 37*, 73–90.

Davys, A. M., & Beddoe, L. (2009). The reflective learning model: Supervision of social work students. *Social Work Education, 28*(8), 919–933.

Dewey, J. (1910). *How we think*. Boston: D.C. Health.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Lexington, MA: Heath.

Dolan, P., Canavan, J., & Pinkerton, J. (2006). *Family support as reflective practice*. London: Jessica Kingsley Publishers.
Finch, J. (1987). The vignette technique in survey research. *Sociology, 21*, 105–114.
Fook, J. (1991). Reflective practice and critical reflection. In J. Lishman (Ed.), *Handbook for practice learning in social work and social care* (3rd ed., pp. 440–454). London: Jessica Kingsley Publishers.
Fook, J., & Gardner, F. (2007). *Practising critical reflection: A resource handbook*. Maidenhead: Open University Press.
Grant, A. M., Franklin, J., & Langford, P. (2002). The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality, 30*(8), 821–836.
Gursansky, D., Quinn, D., & Le Sueur, E. (2010). Authenticity in reflection: Building reflective skills for social work. *Social Work Education, 29*(7), 778–791.
Hatton, N., & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education, 11*(1), 33–49.
Hughes, L., & Pengelly, P. (1997). *Staff supervision in a turbulent environment*. London: Jessica Kingsley.
Ixer, G. (1999). There's no such thing as reflection. *British Journal of Social Work, 29*, 513–527.
Kember, D., Leung, D. Y. P., Jones, A., Loke, A. Y., Mckay, J., & Sinclair, K., et al. (2000). Development of a questionnaire to measure the level of reflective thinking. *Assessment & Evaluation in Higher Education, 25*(4), 381–395.
Larrivee, B. (2000). Transforming teacher practice: Becoming the critically reflective teacher. *Reflective Practice*, *3*, 293–307.
Linacre, J. M. (2012). *A user's guide to Winsteps ministep: Rasch computer programs*. Chicago: Winsteps.com.
Mezirow, J. (1991). *Transformative dimensions of adult learning*. San Francisco: Jossey-Bass.
Nathan, J. (1993). The battered social worker: A psychodynamic contribution to practice, supervision and policy. *Journal of Social Work Practice, 7*(1), 73–80.
Oltedal, S. (2010). Reflections upon practices are important in social work. *Journal of Comparative Social Work, 2*, 1–3.
Regehr, G., Bogo, M., Regehr, C., & Power, R. (2007). Can we build a better mousetrap? *Journal of Social Work Education*, *43*, 327–344.
Reid, B. (1993). "But we're doing it already!" Exploring a response to the concept of reflective practice in order to improve its facilitation. *Nurse Education Today*, *13*, 305–309.
Rogers, R. R. (2000). Reflective thinking in professional practice: A model. *CPD Journal*, *3*, 129–154.
Rogers, R. R. (2001). Reflection in higher education. A concept analysis. *Innovative Higher Education, 26*(1), 37–57.
Ruch, G. (2000). Self and social work: Towards an integrated model of practice. *Journal of Social Work Practice, 14*(2), 99–112.
Ruch, G. (2007). Reflective practice in contemporary child-care social work: The role of containment. *British Journal of Social Work, 37*, 659–680.
Schön, D. (1983). *The reflective practitioner: How professionals think in action*. London: Temple Smith.
Schön, D. (1987). *Educating the reflective practitioner*. San Francisco, CA: Jossey-Bass.
Schön, D. (1993). *Reflective inquiry in social work practice*. Hong Kong: Centre for the Study of Social Work Practice.
Sheppard, J. (2000). Learning from personal experience: Reflections on social work practice with mother and child and family care. *Journal of Social Work Practice, 14*(1), 38–50.
Smith, E. (2011). Teaching critical reflection. *Teaching in Higher Education, 16*(2), 211223.
Social Work Reform Board (2010). *Building a safe and confident future: One year on overarching professional standards for social workers in England*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/180792/DFE-00602-2010-2.pdf
Taylor, C., & White, S. (2001). Knowledge, truth and reflexivity: The problem of judgment in social work. *Journal of Social Work, 1*(1), 37–59.

University of Bedfordshire (2014). *The professional capabilities framework (PCF)*. Retrieved from http://www.beds.ac.uk/howtoapply/departments/appliedsocialstudies/reforms-in-social-work-education/the-professional-capabilities-framework-the-pcf

Wright, B., Linacre, J. M., Gustafsson, J. E., & Martin-Loff, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Yip, K. (2006). Self-reflection in reflective practice: A note of caution. *British Journal of Social Work, 36*(5), 777–788.

Validation of the Pre-licensure Examination for Pre-service Teachers in Professional Education Using Rasch Analysis

Jovelyn Delosa

Introduction

Teachers play a crucial role for students to demonstrate the expected learning outcomes. Teachers are curriculum planners, assessors and curriculum implementers (McTighe and Wiggins 2005). Teachers need to establish a strong nexus between the content that they are giving to their learners, the methods they used and the assessment they employ. One of the roles of teachers is to assess the learning process and outcomes. Tests are widely used as a form of assessment in universities of many countries to measure student learning, to rank students and issue certification. Literature also outlines arguments on how validity issues of tests. There are critics about tests but there are also groups who argued that tests can be used as long as they are reliable and valid. This prompted the researcher to look into her own context specifically in teacher education and examine the test used for the pre-service teachers in preparation for the Licensure Examination for Teachers in the Philippines and look into its psychometric qualities. Using a Mock LET instrument, this paper discusses the strengths of the Rasch model as a psychometric tool and analysis technique, referring to person-item maps and differential item functioning.

The pre-service teachers of Xavier University enroll themselves in a subject called Education 60, a Refresher Course for the Licensure Examination for Teachers. At the end of the course, they are given a test that sets similar to the real Licensure Examination for Teachers (LET). This paper aimed to examine the validity of the test and determine whether this mock test measures the knowledge and skills expected of them. Validation is important to ensure that the given tests are appropriate and the results are trustworthy.

J. Delosa (✉)
Xavier University, Cagayan de Oro, Philippines
e-mail: jingdelosa@gmail.com

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_8

Literature Review

Role of Assessment

Literature confirms the various roles of assessment in learning. Assessments inform all stakeholders about instruction and learning outcomes. Specifically in higher education, assessment is very crucial especially when results of such are the bases of ranking and certification. It drives instruction to important goals and standards (Brew et al. 2009; Rieg and Wilson 2009). Assessment is a component equally important if schools and stakeholders want to increase educational outcomes (Guskey 2003). Assessment is defined as the process of observing, interpreting and making decisions about learning (Griffin 2009). In addition, this process of collecting of evidences can take many forms which include tests, performances, work samples and many others (Black and Wiliam 1998; Griffin 2009). Assessments are given to facilitate learning, facilitate teaching, for school and professional requirements. They provide feedback, motivate students and inform how teachers deliver content to their students and improve their methodologies.

Assessing student performance is one of the most critical responsibilities of classroom teachers (Stiggins as cited in Mertler and Campbell 2005). Assessment information must be correct, reliable, and valid because these information can do a lot to improve classroom instruction (Mertler and Campbell 2005). Teachers are expected to show expertise in assessment (Campbell, Murphy, and Holt as cited in Mertler and Campbell 2005). However, there are issues with assessment. Research has documented that teachers' assessment skills are generally weak (Brookhart; Campbell, Murphy, and Holt as cited in Mertler and Campbell 2005). Among the various roles of teachers, assessment of student learning is somehow left out. Teachers experience inadequacy and difficulty in carrying this role (Murray as cited in Mertler and Campbell 2005), thus, a need to review new frameworks of assessment.

Tests

One of the widely utilized types to assess student outcomes is the use of tests. It could be multiple-item tests, matching type, and true/false tests or fill in the blank. Research shows an important relationship between the quality of classroom assessments and achievement as measured by standardised tests (Mertler and Campbell 2005). Teachers trust the results of tests because of their direct relation to classroom instructional goals and results are immediate and easy to use for analysis since it is still on the student level (Guskey 2003). Webber and Lupart (2012) argued that classroom assessment is the most important kind of assessment.

Multiple-choice items are widely used on classroom tests in colleges and universities (Mavis et al. 2001; McDougall 1997 as cited in DiBattista and Kurzawa 2011). A typical multiple-item test consists of a question and a set of two or more

options that includes the correct answer and distracter options. Multiple-choice tests are commonly used in the classroom and for licensure purposes because grading is easy (DiBattista and Kurzawa 2011) and allows for broader coverage of the topics (Bacon 2003 as cited in DiBattista and Kurzawa 2011). The fact that tests are widely used and inferences are made from the test results raises the challenge for teachers to give valid tests to students to ensure fairness (Popham 2002, 2004 in Webber and Lupart 2012).

Characteristics of a Good Test

A good test should be relevant to the needs of the learners which means that testing is not just an end in itself. It has an educational impact to both learners and teachers and it matches with the curriculum. In constructing tests, teachers should consider the feasibility of the test which includes the time for construction, time for administration, time for scoring and time for reporting (Fuentealba 2011). A good test has validity. It refers to the ability of an instrument to measure the attributes which could be knowledge or skill that it is aiming to measure (Fuentealba 2011; Purya and Nazila 2011). Validity includes looking on the importance of content to me measured, instructions, wording of the questions, spelling and grammar, level of difficulty, arrangement of items, number of items, time and the errors in scoring (Fuentealba 2011). Validity is probably the most important criterion in judging the effectiveness of a measurement tool (Alagumalai and Curtis 2005). Validity has four types: content, predictive, concurrent, and construct validity. Construct validity is the focus of this paper. Construct validity is concerned with the extent to which a test reflects the underlying construct the test is supposed to assess (Purya and Nazila 2011). Valid tests are reliable tests. Reliability is the ability of a test to measure the attributes consistently (Al-Sabbah et al. 2010; Griffin 2009). Reliable assessment when tasks get the same results regardless of when they are administered (Al-Sabbah et al. 2010).

Establishing the psychometric qualities of a test is highly essential and to help teachers we need to look into classical and item response theories. They serve as guiding principles in decision-making of teachers with student learning. Teachers should be equipped with a degree of test literacy which includes test construction test analysis and testing theories. The ability to select and design assessment tools is highly expected of every teacher (Rieg and Wilson 2009).

Classical Test Theory

Test theories provide a framework about the relationship of test and item scores to true scores and ability scores (Hambleton and Jones 1993). Test theories are important in educational measurement because they provide a guiding post for

considering issues of handling measurement errors. “Different models and theories will handle error differently”. One may assume normal distribution of errors and the other may have another assumption (Hambleton and Jones 1993).

Classical test theory introduces three concepts: observed score, true score, and error score. True score is the difference between test score and error score. CTT is a “psychometric theory laid by Charles Spearman in 1904 that allows the prediction of outcomes of testing such as the ability of the test takers and the difficulty of items” (Alagumalai and Curtis 2005, p. 5). This theory explains the concept of an observed score that is manifest and this score is composed of a true score and an error which are both latent in nature. It is a model for testing which is widely used in constructing and evaluating fixed length tests. Although the major focus of CTT is on test-level information, item statistics, like item difficulty and item discrimination, are also important. The p-value, which is the proportion of examinees that answers an item correctly, is used as the index for the item difficulty. A higher value indicates easier items. The item discrimination index is the “correlation coefficient between the scores on the item and the scores on the total test and indicates the extent to which an item discriminates between high ability examinees and low ability examinees”. Similarly, the point-biserial correlation coefficient is the “Pearson r between the dichotomous item variable and the continuous total score variables” (Alagumalai and Curtis 2005, p. 7). However, CTT has limitations (Alagumalai and Curtis 2005). The two statistics which are item difficulty and item discrimination are group and test dependent. The increase or decrease and the homogeneity or heterogeneity of the group affects the results; and test difficulty has a direct effect on test scores (Hambleton as cited in Hambleton and Jones 1993; Boone and Scantlebury 2006). There is no basis to predict how an examinee may perform on a particular item. The true score is not an absolute characteristic of a test taker since it depends on content. A simple or more difficult test would result in different scores for examinees with different levels of ability. Therefore, it is difficult to compare test takers’ results between different tests (Boone and Scantlebury 2006). In the discussion about tests above, testing for many years has a key role in assessing learning (DiBattista and Kurzawa 2011).

However, there are also issues with the use of tests. Guskey (2003) argued that for assessment to be of use, teachers should change their views on how to interpret results. Particularly for multiple-item tests, critics argued that this type of test can be subject to guessing (DiBattista and Kurzawa 2011) and questions on validity and bias (Boone and Scantlebury 2006; Stiggins 1999). Guskey (2003) added that despite of the importance of assessment education today, few teachers receive much formal training in assessment design and analysis.

With this given fact of how tests in universities and countries are used to rank and certify, it is logical to look carefully at the tests that we construct. The purpose of an examination is to infer about students knowledge, skills and values, make inferences about an overt behavior to a covert quality and this poses a problem with scores. There are groups which questioned the reliability of raw scores in representing a person’s true ability.

It is sensible that with the limitations of CTT as cited above, teachers can use another theory, the Item Response Theory (IRT). The past 50 years has not only seen the strengths and limitations of the classical test theory but also acknowledged the use of new approaches to educational measurement.

Psychometricians were interested in psychometric theory which would describe examinees' achievement as independent of the particular choice of items that were used in a test. Classical item statistics such as item difficulty and item discrimination and test statistics such as test reliability are sample dependent; however, thousand of excellent tests have been constructed in this way. Classical test theory and related models have been used and are still used successfully for over 60 years and many testing programs are deeply founded in classical measurement models (Hambleton and Jones 1993).

IRT and Rasch Model

Item response theory is a theory about the relationship of an examinee's performance in an item and with his ability. The items are discreet or continuous and the scores are dichotomous or polytomous. This theory argues that an item can measure single or multiple abilities (Hambleton and Jones 1993).

IRT was originally developed to overcome the issues with CTT. IRT assumes that the latent ability of a test taker is independent of the content of a test. The relationship between the probability of answering an item correctly and the ability of a test taker can be shown in different models depending on the nature of the test. It also assumes that it does not matter which items are used making the possibility to compare test takers (Wiberg 2004).

One of the models in IRT is the Rasch model, named after the Danish mathematician and statistician George Rasch, which is a probabilistic model with two distinguishing properties: invariance (Boone and Scantlebury 2006) and interval scaling which are obtained if unidimensionality occurred that is when the data fit the model (Purya and Nazila 2011). Unidimensionality is achieved when the instrument measures one trait at a time (Wolfe and Smith 2007 in Purya and Nazila 2011).

It specifies the probability of a correct response on an item as a function of the difference between the ability of person and the difficulty of a test item (Webber and Lupart 2012).

When the student ability equals the difficulty of the item, there is a 50 % probability that the student will answer the item correctly (Webber and Lupart 2012). The model is probabilistic based upon logits (Lamb et al. 2011). If data fit the model, the scale is defined as being unidimensional; one can be confident that the item measures are independent of the person measures and vice versa (Purya and Nazila 2011). However if the data do not fit the model, this can be because the

instrument items may be measuring another construct and Rasch analysis allows these items to be identified. When scales are multidimensional, summing of item scores may cause misleading assumptions to be made (Belvedere and de Morton 2010).

Fit indices are used to check the relevance of the test content to the intended construct. Misfitting items may be measuring a totally different and irrelevant construct. Moreover, person-item map and item strata are two important criteria for checking the representativeness of the items (Wolfe and Smith 2007 in Purya and Nazila 2011).

Test takers scores are expressed in logit measures which are the conversion of raw scores to logits through use of the Rasch model. If researchers or educational practitioners do not convert raw scores to equal interval measures then the results of their analysis may provide incorrect and/or incomplete information on student performance. If a researchers uses only raw scores, then incorrect conclusions may be reached by using raw score data for parametric tests of student (Boone and Scantlebury 2006).

Rasch statistics provide similar psychometric information to traditional analyses. A point biserial expresses item discrimination, and a "person separation index". Classical test theory provides a single standard error of measurement (SEM). However, in Rasch measurement each item and test taker is provided an error term. The error in each item is considered and the range of each person's error before item removal. One technique utilized in Rasch measurement is an evaluation of an individual person's responses to test items to the model known as "fit" statistics. Fit statistics are used to assure whether the test is unidimensional and guide one to decide upon the way the test should be scored. However, in case of multidimensionality, separate scores should be reported for each dimension. Thus, fit statistics provide helpful evidence with regard to the structural aspect of construct validity (Beglar 2010; Purya and Nazila 2011). Item fit statistics evaluate the predictability of test takers' answers, given their overall ability (Boone and Scantlebury 2006).

Differential Item Functioning

Rasch analysis also facilitates the assessment of differential item functioning (DIF). DIF occurs when persons of the same ability have items that operate differently based on another variable, such as age or gender. Assessment of DIF is important as it improves generalisability of the instrument by testing that item response patterns are similar across groups. Rasch analysis also facilitates the investigation of item thresholds. If the probability of each item response category is not in the expected order, this results in a disordered threshold (Belvedere and de Morton 2010).

Hambleton and Jones (1993) summarized the main differences between classical test theory and item response theory. Classical test theory is linear, the level is for the whole test, assumptions are easy to meet test data, item-ability relationship is not specified, test scores or estimated true scores are reported on the test-score scale,

item and person parameters are sample dependent, while on the other hand, item response theory is nonlinear, level of analysis is by item, assumptions are difficult to meet with test data, item characteristics functions are available, ability scores are reported on a transformed scale, item and person parameters are sample independent if model fits the data. CTT is test based while IRT is item based. CTT permits no consideration of how participants respond to a specific item. IRT permits the analysis of the probability of an examinee answering an item. In item response theory; the measurement specialist is allowed a greater flexibility.

IRT has limitations too because of its complex mathematical formulations; however, with technology nowadays it is now very possible for teachers to analyze tests using software. Another thing is if a test is poorly designed, computing an overall measure using all test items may be impossible and results may only be evaluated at the item level. A Rasch analysis may take longer than a traditional analysis, but it provides a deeper understanding of instrument's strengths and weaknesses (Boone and Scantlebury 2006).

Licensure Tests

After the comparison between CTT and IRT, let us take a look at a specific kind of test, licensure tests. Many countries including the Philippines have practiced the national examination for licensing of teachers before teachers are considered professional teachers. The role of teachers in education has been identified as the most significant of all school factors that affect student learning and with this belief, policymakers want to guarantee a level of quality through a licensure system. Teachers pass licensure tests given by the government before they can work in the classroom. Licensure is defined by the US Department of Health "as a process by which an agency of government grants permission to an individual to engage in a given occupation upon finding that the applicant has attained the minimal degree of competency required to ensure that the public health, safety, and welfare will be reasonably well protected" (Shimberg 1981). Teacher preparation programs are being held to high standards in order to prepare the best teachers to meet the challenges of today's diverse classrooms (Rieg and Wilson 2009). The main mission of teacher education in the Philippines is the training and preparation of globally competitive teachers who are equipped with the principles, aspirations and values and possess pedagogical knowledge and skills (CMO 30, 2004). With this goal, teacher education institutions are challenged to develop and guide pre-service teachers towards this direction. To professionalize the teachers, RA No. 7836, known as Professionalization Act for Teachers is implemented to "strengthen, regulate and supervise the practice of teaching profession in the Philippines by prescribing a license" to teachers certified by the Professional Regulation Commission (PRC). The Professional Regulation Commission works hand in hand with the Commission in Higher Education (CHED) with the Teacher Education Institutions (TEI's) in the implementation of this law. CHED issued CHED

Memorandum 30, s. 2004 (CMO 30, 2004) which provided a list of the desired competencies and subject areas to be taken by pre-service teachers. The list included General education Courses (Science, Math, English, etc.); Professional courses which have three subgroups, theory subjects, strategies subjects and field subjects; and Specialization courses. PRC issued a list of competencies based on the National Competency-Based Standards (NCBTS) and their weights. For elementary education (BEED), 40 % is allotted for general education and 60 % for professional education; general education (20 %), professional education (40 %) and specialization (40 %). The TEI's created partnership with the Department of Education in coming up with the Experiential Learning course for the pre-service teachers which provided students with actual learning experiences.

Examining the PLET Items

Pre-licensure Examination for Teachers Test (PLET)

This section discusses the background of the practice test used before the teachers take the real Licensure exam for teachers. This particular study focused on the professional courses for LET. These items are constructed by the different teachers teaching the subjects. The LET coordinator compiled all the items and gives the test as the final examination for the subject Education 60. It is a 6-unit course. The class runs for 14–15 Saturdays and the sessions from 1–7 pm. The sessions cover all the competencies indicated in the LET Primer and address the three components of General Education, Professional Education and Major area of concentration.

To demonstrate content validity, it is important to establish that the questions on a test represent a content domain that the test sample (Shimberg 1981). The questions of this test are based on the list of competencies as seen in the LET Primer. When the researcher reviewed the items, 90 % of the items are reflected in the competencies in the LET Primer. However, nothing much has been done to examine the construct validity of this test except that of looking at the overall scores of the students and checking what items were not answered right. This then is the main purpose of this paper, to investigate the one construct validity of the test using the IRT perspective.

Item Analysis Using the Rasch Model

The Rasch model is used for item analysis. It has features which Rasch labeled as 'specific objectivity' and unidimensionality. Unidimensionality is when the items measure the same construct while specific objectivity states that two person who are taking the tests are compared and such comparison is not based on that items are included in the test. Item analysis using Rasch model can give practitioners the

answers why particular tests are not functioning as they should and can guide them on which items to include or to omit (Choppin 1983). Validity was assessed by evaluating the fit of individual items to the latent trait as per the Rasch model and examining if the pattern of item difficulties was consistent with the model expectancies.

The Data

The test is composed of 200 items. These items tried to measure the students' understanding of professional education which is one of the major components in the real Licensure Examination for Teachers. This test investigated their knowledge about theories, assessment skills and other pedagogical areas of the teaching profession The 200 items were subjected to Rasch analysis to check their fit and if all these items measure one construct which is professional education. Fit indices were examined closely to check the relevance of the items as part of content validity. However, the results of the analysis showed that the 200 items are composed of other constructs because of the variety of patterns of responses no matter how many times the test was rerun in Conquest. The researcher decided to check the items again. There are 7 constructs that emerged from the 200 items based on the content analysis of the whole test vis a vis the list of teaching competencies: understanding of curriculum concepts, understanding of teaching profession concepts, knowledge and application of educational technology concepts, understanding of social dimension concepts, knowledge and application of assessment concepts, application of teaching principles, and understanding of the theories of learning. Curriculum items summarise topics about the nature of curriculum development and there are 6 items. The teaching profession domain is about knowledge on the laws that govern the teaching profession and the essence of the profession; it has 7 items. Educational technology items include competencies on the use of technology in the classroom with 14 items. Social dimension concepts include understanding of the role of society in education with 23 items. The assessment construct which describes various concepts of assessment knowledge and skills have 41 items; principles of teaching which is about the strategies in teaching has 47 items and theories of learning has 62 items. The last domains have more items compare with the rest since these topics belong to subjects which are credited for 6 units in the entire pre-service education.

Findings

Item Analysis of the Curriculum Items

The curriculum items were subjected to Rasch analysis using the residual-based fit statistics. The important information considered were the Infit Weighted Mean Square (IWMS) and the t-statistic (T) which determined whether an item followed

the requirements of measurement. A range 0.80–1.20 (Wright and Linacre 1994) was used for IWMS since LET is a test, and −2 to +2 for calculated T (Wu and Adams 2007) to indicate acceptable mean fit. The items with mean square values falling above 1.2 were considered under fitting and suggests the presence of unexpectedly high variability (Bond and Fox 2007 in Franchignoni et al. 2011) and do not discriminate the students with high ability from those with low ability while values below 0.8 were over fitting items and gave redundant information and too predictable pattern (Wright and Linacre 1994). The items that do not fit the model were removed one at a time. In this paper, only the initial and final analyses results are presented. The reliability was evaluated in terms of 'separation', defined as the ratio of the true spread of the measures with their measurement error (Franchignoni et al. 2011).

The initial analysis included all the curriculum items and 152 test takers. The results are tabulated in Appendix A. In the first run of the analysis, all items belong to the 'good' fit so there was no need to rerun it. The item difficulty values are shown in Table 1 and they are expressed in terms of logit, the unit used in Rasch logit interval scale which allows person and item to be placed on a common scale (Wright and Linacre 1994). The scale consists of numbers from $-\infty$ to $+\infty$ with 0 in the middle indicating average difficulty for item and person (Bond and Fox 2007). Items with estimates above 0 (positive values) are more difficult items and those below 0 (negative values) are easier items. However, the items level of difficulty and arrangement is needed to be revisited because the items were not arranged well. Item 1 was easy then item 2 was very difficult. Additionally, items in this construct showed good discrimination ability.

The Rasch model transforms raw item difficulties and raw person scores to equal interval measures. These measures are used to map persons and items onto a linear scale. Items ranged from easiest which is located at the base of the graph and to hardest located at the top of the graph. Persons are plotted as a function of their ability, with the more able students at the top of the graph, and less able students at the base. Items plotted above any person are harder than the person's ability level and items below a person are those items for which the person has a greater than 50/50 chance of correctly answering (Santelices and Wilson 2012).

Item Analysis of the Teaching Profession Items

The 7 items of teaching profession were subjected to Rasch analysis using the residual-based statistics. Their IWMS and T-values were examined for fit to measurement requirements. The complete analysis is reported in Appendix B. The infit weighted mean square was used to identify the rating of the items that deviate from expectations (Wright and Linacre 1994). All items were in 'good' fit with IWMS falling within the range of 0.80 to 1.20. This is the final results since all items conformed to the measurement requirement. Table 2 presents the results of the analysis showing that all items belong to the desired IWMS and T-values. Similar

to the curriculum items, the arrangement of items according to the level of difficulty needed some attention. Item 3 is the most difficult item; and item 6, the easiest. Majority of the students were in the average level of ability.

Item Analysis of the Educational Technology Items

The results revealed that in the initial analysis all items achieved the required IWMS and T-values. The separation reliability which is defined as the ratio of the true spread of the measures with their measurement error (Bond and Fox 2007, pp. 40–41 as cited in Franchignoni et al. 2011) is high (0.986). It can be observed that the average ability of the students is above the difficulty of the items. Item 3 is the most difficult item and item 6, the easiest. Majority of the students were in the average level of ability. It is essential to review the difficulty level of the items because items 3, 13, and 14 are too easy and far below the ability of the students and items 1 and 6 are too difficult for the students.

Item Analysis of the Social Dimension Items

There are 23 items for social dimension. The results showed that in the initial analysis, there was 1 underfitting item which was removed and the analysis was rerun. This process was done for the second time. In the second analysis, all the 22 items conformed to the fit requirement. All the items have IWMS and T-values that are within the range of the required measurement values. Table 3 shows the second and final analysis of the items, their particular estimates, error, IWMS and T-values. The social dimension items have a separation reliability of 0.988 which conformed to the required value of equals to and more than 0.90 (Wright and Linacre 1994). Checking at the estimates, the items were not arranged well from easy to difficult. Item 2 which was removed in the second run has a T-value of 2.2 which means that this is an under fit item (Wu and Adams 2007). Half of the items are difficult items with estimates of positive values. Most of the items have a discrimination value which was quite good except for item 11 which has a negative value which means that this item did not discriminate the students with high ability from the students with low ability. Even if the items have reached the required measurement fit it is important to reexamine the structure of the questions to improve them.

Item Analysis of the Assessment Items

The initial analysis included 41 items which were examined for measurement fit. The item with the ‘worst’ fit (based on the T-value and IWMS) was first removed

and the analysis was done again. This step was performed until no misfitting items were found. One item was found to be over fitting and two to be under fitting based on their T-values. The under fitting item was first removed and the analysis was rerun. This process was done until all items fit the measurement requirement. Two more analyses were done to come up with 39 'good items' whose IWMS and t-statistics fell under the required range of measurement. Item 26 was the most difficult item that no one got it right and items 10 and 35 were the easiest that all students got them right.

Item Analysis of the Principle of Teaching Items

Similar with the other constructs, the 47 items of the principle of teaching were subjected to the Rasch analysis for dichotomous items. In the initial analysis, two under fitting items were found. The item with the biggest T-value was removed first and the data was rerun. The second analysis showed that all the 46 items were 'good' items as based on their infit weighted mean square and T-value. Forty-five percent of the items or 21 items out of 46 items were difficult items and the separation reliability is 0.987. Some items need to be reviewed because they are too difficult (PT 24, 35, 36, 43) and too easy (PT 4, 15, 17, 25 and 37).

Item Analysis of the Theories of Learning Items

The 62 items of the theories of learning were subjected to the Rasch analysis for dichotomous items. In the initial analysis, three under fitting items were found. The item with the biggest T-value was removed first and the data was rerun. The third analysis showed that all the 59 tems were 'good' items as based on their infit weighted mean square and T-value. The separation reliability is 0.987. Most of the items were also below the average level of difficulty. There were also outliers, items which were extremely difficult that no student got the right and extremely easy that even students with ability below average got them right.

Differential Item Functioning (DIF) of the Pre-licensure Examination for Teachers (PLET) Items

The 7 constructs of the 200 item pre-Licensure test were submitted to Differential Item Functioning (DIF) after examining that the items are good items. DIF was done to examine if the items behave well across the two groups: male and female in this particular study. DIF is necessary to check whether test items have same response patterns to ensure generalizability (Boone and Scantlebury 2006).

All items from the 7 constructs which have acceptable fit were analyzed. There were 6 items for curriculum, 14 items for educational technology, 7 items for teaching profession, 22 items for social dimension, 38 items for assessment, 46 items for principle of teaching and 59 items for theories of learning. Items of each construct were examined separately using Conquest 2.0 software (Wu et al. 2007). In DIF detection, these indicators were considered: an approximate Z-statistic (calculated T-value) calculated by dividing the estimate by the standard error; comparing the standard error with the parameter estimate (Wu et al. 2007); checking if the chi-square value is significant (Wu et al. 2007) and verifying the difference between the estimates. A calculated T-value less than −2.0 or greater than +2.0 points out significant DIF between two groups and in this test chi-square value of equal to and less than 0.05 is significant.

The 7 constructs of the 200 item pre-Licensure test were submitted to Differential Item Functioning (DIF) after examining that the items are good items. DIF was done to examine if the items behave well across the two groups: male and female in this particular study. DIF is necessary to check whether test items have same response patterns to ensure generalisability (Boone and Scantlebury 2006). All items from the 7 constructs which have acceptable fit were analyzed. There were 6 items for curriculum, 14 items for educational technology, 7 items for teaching profession, 22 items for social dimension, 38 items for assessment, 46 items for principle of teaching and 59 items for theories of learning. Items of each construct were examined separately using Conquest 2.0 software (Wu et al. 2007). In DIF detection, these indicators were considered: an approximate Z-statistic (calculated T-value) calculated by dividing the estimate by the standard error; comparing the standard error with the parameter estimate (Wu et al. 2007); checking if the chi-square value is significant (Wu et al. 2007) and verifying the difference between the estimates. A calculated T-value less than −2.0 or greater than +2.0 points out significant DIF between two groups and in this test chi-square value of equal to and less than 0.05 is significant.

DIF in Curriculum Items

The overall results of DIF analysis of the curriculum items by gender shows that the LET curriculum items exhibited no DIF as evident in its T-value of 1.48 and the parameter estimate is lower than twice its standard error. The results also show that on average female pre-service teachers perform higher the males with a logit difference of 0.148 but this difference is not significant (chi-square p value = 0.141; calculated T-value = −1.48).

The item level results indicate that females achieved higher in 3 items and in the same manner, males achieve higher in 3 items; however, the difference is not statistically significant. Curriculum items behaved the same between the two groups.

The result shows a minimal difference of 0.296 of performance between males and females yet it is to be taken into account that this difference is not significant.

DIF in Teaching Profession Items

The results show that the LET teaching profession items exhibited DIF based on the calculated T-value which is more than −2 (T-value = −4.98). The female pre-service teachers on average scored higher than the male pre-service teachers in the knowledge and application of concepts about the teaching profession with a difference of 1.006 logit which indicates a gap of 2 years (Griffin 2009). The parameter estimate is more than twice its standard error and the fact that the chi-square value is smaller than 0.05 (significant level = 0.001) indicate that this difference is statistically significant. These results have some implications on revisiting the items about the teaching profession. These items try to measure understanding about teaching as a mission, a vocation and profession. The items should function fairly to both groups to establish fairness even if in the Philippines, the teaching profession is mostly embraced by women.

However, at the item level, the results varied slightly. Based on the calculated T and the estimate compared with twice of the standard error, there is only one item that shows DIF and the rest of the items fall along the range of −2 to + 2 for calculated T. Out of the 7 items, males performed higher than females in 4 items. Nevertheless, the results strongly suggest that the items behaved differently between the two groups based on gender and therefore, it is important that these items need to be reviewed. The result shows that the overall teaching profession items manifest an achievement difference of 1.006 between males and females with the female teachers outperforming the males.

DIF in Educational Technology Items

DIF is not evident in the LET educational technology items. The overall results reveal that females on average got a higher achievement than the males with a logit difference of 0.20 which is minimal. DIF analysis in the item level demonstrates a degree of varied item responses by gender. There are 14 items in this construct and females performed higher in 7 items as well as male did. In determining for DIF in the item level, some items showed no DIF as based on the Z-statistic (calculated T-value) and the result of comparing the estimate with twice of the item's standard error like items 4, 6, and 7; however, there is a need to review these items because even if the big difference between estimates even if the difference is not statistically

significant. The test makers of this item can improve the word structure of the item or syntax of the item. The result validates that DIF did not exist in the educational technology items. With the sweeping influence of technology in education, people are trying to cope with these changes, both males and females.

DIF in Social Dimension Items

DIF is not also present in the LET educational technology items. The overall results as shown in Table 11 reveals that females on average got a higher achievement than the males with a logit difference of 0.23; however, their difference between their estimates was not statistically significant. Items in social dimension construct tried to ask students their understanding of the society and the function of school in the different system that governs human activities.

The Millennium Development Goals of the Philippines (MDG) advocated for gender equity as both males and females tried to portray their roles as teachers. The notion that teaching is regarded as a women work has a long history (Skelton 2002). There is a need of male teachers to serve as good role models for males. Schools have the responsibility to ensure gender equity between males and females. Gender equity should be fundamental in educational practices. DIF analysis in the item level demonstrates a degree of varied item responses by gender. There are 14 items where females performed higher and 8 items where the males achieved higher. In determining for DIF in the item level, some items showed no DIF as based on the Z-statistic (calculated T-value) and the result of comparing the estimate with twice of the item's standard error like items 3, 4, and 8, 9, 11, 13, 19, 22; however, there is a need to review these items because even if the big difference between estimates even if the difference is not statistically significant. The test makers of this item can improve the word structure of the item or syntax of the item or the arrangement of the items. It is recommended to recheck the items. The result validates that DIF did not exist in the educational technology items. The females on average performed better than the males. Even DIF did not significantly exist, it is interesting to observe that in item 8, males performed much better than the females who were even below the average ability level but in most of the items, both groups gathered in the middle near the average ability and difficulty.

DIF in Assessment Items

The overall DIF analysis for the assessment items show that the items function the same across the 2 groups ($T = -0.65$). Results revealed that females on average got a higher achievement than the males with a logit difference of 0.26; however, their

difference between their estimates was not statistically significant. Assessment items tried to measure knowledge on both traditional and alternative forms of assessment. Every teacher is expected to acquire the skills on how to assess learning in two methods: the use of tests and authentic assessment. The pre-service teachers of the university were constantly exposed to avenues to hone their assessment knowledge and skills. The results showed that the ability of the test takers is higher than the average ability.

DIF analysis of assessment items in the item level demonstrates that females got a higher achievement in items in 17 items while males achieved better in 29 items, however, in the overall analysis, the females did well than the males in the test. It is also noted that even in the overall analysis, DIF did not exist, there are items that should be reviewed because of the big difference between estimates of males and females though the difference is not significant. The results revealed that items 26, 14 and 24 were items too difficult for the test takers to answer and items 23, 1, 35, 10 and 37 were too easy that everyone got it right. Similarly with teaching profession items, principles of teaching items showed a significant DIF. The content analysis of the topics for these two areas showed that many topics were related much with each other. Again, it was the group of the female which achieved higher than males with a difference of nearly 0.5 logit which is equivalent to 1 year of learning.

DIF in Principle of Teaching Items

The item level analysis presented item 32 showing DIF with a T-value of −2.03 and a difference of 0.89 logits between males and female with females outscoring the males. There are also 18 items which needed some reexamination even if there is no significant difference according to gender due to big difference in the estimates. The principles of teaching items assessed the students' knowledge and application of classroom strategies and methods of teaching. The results clearly confirmed that the ability of the female test takers was above the average ability and the males' ability was below the average level of ability and this difference is statistically significant (sig level = 0.000).

DIF in Theories of Learning Items

The overall DIF analysis for the theories of learning shows that there is no error in the estimates and therefore T cannot be calculated. However, based on the chi-square test of parameter equality = 0.00, df = 1, and Sig Level = 1.000, it shows that there is no significant difference between the two groups. The separation reliability is 0.96 but the 0 chi-sqaure signaled for a careful reexamination of the

items. The value of chi-square which is 0 could be traced to the homogeneity of the group.

The item level analysis of DIF confirms the overall results showed some conflicting results from the overall results because there are two items which shoed DIF on the item level and many items which needed revalidation. There were items whose T-value was within the range of −2 to +2; however, their difference of estimate is quite big and even if the difference is not significant, it warrants some attention.

Conclusion

In summary, out of the 7 constructs, generally, on average the female achieved higher than the males in all the constructs. All the constructs got separation reliability above 0.95.

Females performed better than the male teachers; this result is to be confirmed yet as to who achieves higher in LET in the Philippines, males of females since only one study has been found investigating gender differences in LET performance.

With regards to DIF, teaching profession and principles of teaching items exhibited significant DIF.

In conclusion, each of the 7 construct measures what it is supposed to measure and each construct has good psychometric qualities. Three of the constructs (curriculum, educational technology and teaching profession) have all items which have the required fit; social dimension has only 1 item which was under fitting; assessment has 2 'worst' items, social dimension has 1 under fitting item, principle of teaching having 1 under fit item and theories of education with 1 over fitting item. The separation of all the constructs is good. Items of each dimension possess a good discrimination power of separating the student which has a high ability from the less performing students. However, putting all the items together to measure the general construct on understanding and application of professional education created a problem thus, there is a need to review the competencies in each construct vis-avis the general goal of professional education as a whole. This maybe because of the large number of items which is 200 and in reality it is difficult that all items measure one construct since the items can be related to one another, there could be overlapping of content. This clearly calls for attention especially when 2 of the constructs exhibited DIF. One important aspect that needs much time for re examination is the development of the assessment instrument. Hence, the following recommendations are hereby given.

Recommendations

The LET items can be improved, first by construct then as a whole measuring professional education which is a major component in the BEED and BSED Licensure Examination for Teachers. Based on the results of the literature review on the role of assessment and tests; and the benefits of IRT with the analyses of the test instrument following recommendations are suggested for future practice.

A great need to create an assessment committee in the School responsible for studying the validity of the test specifically content and construct validity. The team has to invite PRC and CHED to discuss the competencies for LET. Teacher just read the LET Primer and each one could have his/her own understanding of the stated competencies. It would be better that everyone has a common understanding of the content. This collaboration is also important because the subjects are not just taught by one teacher; content should be brought to a common ground too.

- The teachers developing the items have to establish first reliability of the items, do pilot testing of the items before these items be included in the final pool of items for Educ 60 and later develop an item bank.
- A review of the LET Primer whether the school is teaching all the competencies in the Primer since even in the National Licensure examination for Teacher, Professional Education is one area which acquired low performance from test takers.
- Review the level of difficulty of the items in terms of order and arrangement in the test. Many of the items did not follow a pattern of order of simpler to most difficult as they are presented in the test and the construct with DIF.
- Items which are under fitting or over fitting need to be considered for revision. Review also the items which were too difficult that no one answered them and some even skipped them and the easiest items where all the participants were highly able to answer them correctly because the difficulty of these items were far below their ability level. Revisit the theories of learning items and do further investigation of the emergence of a 0.00 chi-square and a significance level of 1.00, thus producing no error.
- Time element is test administration should be reconsidered if the test was administered the same way during the real LET.
- The mock LET instrument should be designed as following the format of the real LET like having booklets where students do the shading of their correct answer and not writing the letter of their choice. This will improve face validity of the test and consistency of the instrument.
- Feedback should be sought from the School of Education alumni who have taken the National Licensure Examination for Teachers on content and procedures. It is not a matter of asking them what came out during the test; it is a matter of soliciting feedback if the subject matter taught in the pre-service are consistent with the knowledge sought during the national test.
- Investigate the consistency of the number of items in the PLET to the items given in the national test.

- Conduct a study about Let exams by collaborating with PRC. There should be bases on LET performance because there is a big gap in between evidenced-based practice and reality. PRC and teacher education institution are rich terms of data about teacher performance in national examinations yet there are no researcher available for reference.
- Incorporate IRT in assessment courses so that pre-service teachers will also be aware of this test theory and its application not just because of LET per se but for their future assessment roles in the classroom.
- Additional research should be conducted to better understand the dynamics of student views for teaching profession and principles of teaching related topics, activities, and pedagogical approaches. Of particular importance is an understanding of the factors that are most important for female and male students since these are two areas where DIF occurred.
- Teacher's teaching profession and principle of teaching topics should intensify the use of research results of gender based studies to design and develop standards based activities that appeal to males.
- The School as a center of excellence shall venture into studying new theories to testing which are now widely used which is the Item response theory, be open to changes in measurement and assessment community. The school should reexamine consistency between content, assessment and methods (Martone and Sireci 2009).

Generally, the pre-Let instrument has good measurement properties. It is an acceptable tool but the items can just be used for each construct only and not to measure the general construct which is professional education. There are only few 'misfit items' per construct however there is an urgent need to have a committee to revisit the items as a whole test for professional education. There seems a problem with unidimensionality if all the items are taken as one test. Each construct is fine but bringing all the constructs together creates some questions on validity. It could be that some items are highly dependent on other items. Another important reminder for teacher educators and other stakeholders that a good licensure test or a pre-licensure test will not remove the need of teacher evaluation and other programs to ensure competence in the teaching field (Mehrens 1987) because there is still that question of whether licensure exams can assure schools for quality teaching (Shepard 1991). CTT continues to be an important framework for test construction. Teachers need to have a clear idea of the relations between IRT and CTT. This should improve the appreciation of both theories and facilitate communication with researchers, item writers and classroom teachers who are frequently more familiar with CTT than with IRT (Bechger et al. 2003).

This study opens new directions for further research in test development of pre-licensure exam of teacher education institutions which can be used not only for this particular school but for the TEI's in the region and even in the country. Hence, it is recommended that an assessment committee in the School of Education should be created; teachers do pilot testing of the items before these items be included in

the final pool of items for Educ 60 and later develop an item bank; review of the LET Primer; review the level of difficulty of the items in terms of order and arrangement in the test; and review of the items which were too difficult that no one answered them and some even skipped them and the easiest items where all the participants were highly able to answer them correctly because the difficulty of these items were far below their ability level.

References

Al-Sabbah, S. A., Lan, O. S., & Mey, S. C. (2010). The using of Rasch-Model in validating the Arabic version of multiple intelligence development assessment scale (MIDAS). [Report]. *International Journal of Behavioral, Cognitive, Education and Psychological Sciences, 2*(3), 152.

Alagumalai, S., & Curtis, D. D. (2005). Classical Test Theory. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 1–14). The Netherlands: Springer.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, *27*(5), 319–334.

Belvedere, S. L., & de Morton, N. A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *Journal of Clinical Epidemiology, 63*(12), 1287–1297.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, N.J: Lawrence Erlbaum Associates Publishers.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*, 253–269.

Brew, C., Riley, P., & Walta, C. (2009). Education students and their teachers: Comparing views on participative assessment practices. *Assessment & Evaluation in Higher Education*, 1–16.

Choppin, B. (1983). *The Rasch model for item analysis*. Center for the Study of Evaluation, University of California.

DiBattista, D., & Kurzawa, L. (2011). Examination of the Quality of Multiple-Choice Items on Classroom Tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2*(2).

Franchignoni, F., Ferriero, G., Giordano, A., Sartorio, F., Vercelli, S., & Brigatti, E. (2011). Psychometric properties of QuickDASH—A classical test theory and Rasch analysis study. *Manual Therapy, 16*(2), 177–182.

Fuentealba, C. (2011). The role of assessment in the student learning process. *Journal of Veterinary Medical Education, 38*(2), 157.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Instructional Topics in Educational Measurement,* Module 16. Retrieved 30 November 2011 from http://www.ncme.org/pubs/items/24.pdf.

Griffin, P. (2009). Teachers' use of assessment data *Educational assessment in the 21st century* (pp. 183–208). Springer.

Guskey, T. R. (2003). How classroom assessments improve learning. *Educational Leadership, 60* (5), 6–11.

Lamb, R., Annetta, L., Meldrum, J., Vallett, D., Lamb, R., Annetta, L., & Vallett, D. (2011). Measuring science interest: Rasch validation of the science interest survey. *International Journal of Science and Mathematics Education, 10*(3), 643–668.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332–1361.

McTighe, J., & Wiggins, G. (2005). *Understanding by Design* (Expanded Second Edition). Association for Supervision & Curriculum Development.

Mehrens, W. (1987). Validity issues in teacher licensure tests. *Journal of Personnel Evaluation in Education, 1*(2), 195–229.

Mertler, C. A., & Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the Assessment Literacy Inventory.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, QC.

Purya, B., & Nazila, A. (2011). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2*(5), 1052.

Rieg, S. A., & Wilson, B. A. (2009). An investigation of the instructional pedagogy and assessment strategies used by teacher educators in two universities within a state system of higher education. *Education, 130*(2), 277–294.

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement, 72*(1), 5–36.

Shepard, L. A. (1991). Will national tests improve student learning? *The Phi Delta Kappan, 73*(3), 232–238.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36*(10), 1138–1146.

Skelton, C. (2002). The 'feminisation of schooling' or 're-masculinising' primary education? *International Studies in Sociology of Education, 12*(1), 77–96. doi:10.1080/09620210200200084.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*(1), 23–27.

Webber, C. F. E., & Lupart, J. L. E. (2012). *Leading student assessment* (Vol. 15). Dordrecht: Springer, Netherlands, Dordrecht.

Wiberg, M. (2004). Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*: Educational Measurement Solutions Melbourne.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ConQuest Version 2.0. Camberwell, Victoria: ACER Press.

Wright & Linacre, (1994) http://www.rasch-analysis.com/rasch-analysis.htm.

Accessing Gender Bias in Malaysian Secondary School Students' Leadership Inventory (M3SLI)

Mei-Teng Ling and Vincent Pang

Background

In order to progress with the changing world, as well as to prepare for the future, the Ministry of Education (KPM) has introduced a new plan in 2012 called 'The Education Development Plan' (PPPM). This plan outlines the directions for the education system in Malaysia for 2013 until 2025. Through this plan, students are supposed to acquire six main attributes, namely thinking skills, knowledge, national identity, ethics and spirituality, bilingual proficiency and leadership skills (Ministry of Education Malaysia 2012).

According to PPPM, the education system helps every student reach his/her potential by creating formal and informal chances for the students to work in teams and to take on leadership roles. In PPPM, leadership encompasses four dimensions; namely, entrepreneurship, resilience, emotional intelligence and strong communication skills. Entrepreneurs take the initiatives to create and develop their own solutions are willing to invest their own resources and have the drive to see these through to the realisation of their aims (Ministry of Education Malaysia 2012). Resilient persons are able to develop constructive idea and are able to overcome obstacles. Emotionally stable persons have the ability to understand and work effectively with others. In addition, they can influence others positively. People who have strengths in communication skills have the ability to express their views and intentions clearly in oral and written form.

It is important to develop students' leadership and 'leadership identity' or sense of self to be able to lead (Renn and Ozaki 2010) during their school life to become

M.-T. Ling (✉) · V. Pang
Faculty of Psychology and Education, Universiti Malaysia Sabah,
Kota Kinabalu, Sabah, Malaysia
e-mail: lingmeiteng@gmail.com

V. Pang
e-mail: pvincent@ums.edu.my

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_9

future leaders. Amirianzadeh et al. (2010) view that leadership is considered as a part of lifelong learning and multidimensional construct involving competency, experiences and process. All adolescents possess leadership potential; 'today's young people are ultimately tomorrow's leaders' (Fertman and Van Linden 1999: 16). Furthermore, Bisland (2004) states that leadership development should begin as early as preschool or kindergarten. As students become older, leadership development should include exposure to and interaction with adult leaders in the community and region through mentorship and internships (Hine 2011). Leadership provides opportunities, stimulate motivation, and create goals to build capacity to promote the use of existing resources and the development of additional resources (Minckler 2011).

Intelligence, aptitude tests and school grades cannot necessarily predict the success of education and life. Instead, skills such as communication skills, patience and goal setting, should be taken into consideration (McClelland 1973). These skills can be developed through leadership development. The concept of leadership in this study refers to the competencies associated with personality and values as a student leader. Therefore, in leadership, personality development and self-esteem for a student should be emphasised. Some schools, especially boarding schools, give little emphasis on personal development and leadership coaching. The majority that provides these aspects do not pay attention in depth (Hassan and Safar 2010). Furthermore, student leadership is not specifically evaluated in schools in Malaysia.

In order to develop and plan a better training module, the schools need to know which specific areas to focus into. Good student leadership development programmes assure that the students not only gain optimum benefits from the training provided but also avoid unnecessary wastage of time and resources (Zakaria et al. 2008). As an effort to determine the student leadership competency profile, the purpose of this study is to develop a valid and reliable instrument to measure student leadership competency. This can help in the planning of programmes for intervention and improvement on students' leadership competency before the students enter higher level education or job market.

Besides, studies on the construction and validity of leadership instruments in Malaysia mostly lacked focus on items that may be bias (Mustamin 2013; Othman et al. 2014; Rosseni et al. 2009). Only some of the researcher were concerned about the biasness of the items (Haslina 2013; Zamru and Anisah 2012; Rahayah et al. 2010). To date, studies on biasness in instruments are less carried out in Malaysia.

Many previous research show differences in leadership role and leadership competency between male and female (Connerley et al. 2008; Farver 2007; Posner 2012). Paustian-underdahl et al. (2014) found that females were rated as significantly more effective than males. In different point of view, Johnson et al. (2008) argue that female leaders will be seen as effective when they show sensitivity and strength while male leaders just need to show strength. Connerley et al. (2008) suggest that although female leaders can be seen as having the same level of performance with male leaders, females are not seen as prepared as men leader at the same rate in international duties. However, Ismail et al. (2013) assert that leadership is not owned by a particular person or a particular gender.

Every instrument that is set to measure students' competency level especially those from multiple background and characteristics must be fair and equal (Rahayah et al. 2010). As an effort to ensure that the instrument is built for male and female students equally, gender differential item functioning (GDIF) is used to detect the possibility of any biased item. This study aims to evaluate the potential of measuring biasness between genders in the M3SLI. The potential biased items were investigated by conducting DIF analysis in Rasch.

Methodology

A quantitative approach that involves the collection of data using survey was applied in this research. The 68-item Malaysian Secondary School Students' Leadership Inventory (M3SLI) based on Tubb's Model (Tubbs and Schulz 2006) was employed to gauge the three domains which comprises of personality (15 items), values (18 items) and competencies (35 items). The instrument was administered to 2340 students from 26 schools in four main divisions of Sabah. The respondents were all from government secondary schools within the state.

While analysing DIF, Bond and Fox Steps was used to perform two-tailed t-test to test the significant difference between two difficulty indices. Confidence level of significant was chosen at 95 % and the critical t value was set at ±2.0 for all DIF analysis. Besides, DIF contrast was also used to determine the difference between item difficulties for the two groups. The size of DIF which is less than 0.5 logit or more than −0.5 logit is negligible. DIF statistical significance is influenced by the size of DIF effect and the size of the classification groups but it is uninfluenced by the model fit (Linacre and Wright 2012). The indicators of DIF used were (1) t value ± 2.0 ($t \geq +2.0 \leq -2.0$), (2) DIF contrast ±0.5 (DIF Contrast $\geq +0.5 \leq -0.5$) and (3) $p < 0.05$ (Bond and Fox 2007).

Psychometric Properties of Instrument

The reliability indices for personality, values and competencies are 1.00, 1.00 and 0.99, respectively, which is strong and acceptable because they are more than 0.8. The high item reliability might be due to the wide difficulty range of items and a large sample size. When the index is high, the sample size is enough for stable comparisons between items (Linacre and Wright 2012). The item separation index of M3SLI is 15.59, which is acceptable, based on Linacre and Wright (2012) that the separation index of more than 2 is good. High item separation (>3, item reliability > 0.9) implied that the person sample is enough to confirm the item difficulty hierarchy, which is the construct validity of the instrument. The higher the number of separations, the more confidence the researcher can replicate the items across other samples (Linacre and Wright 2012). Besides, positive PTMEA Corr. indicates

that all the items are functioning towards a single measurement constructs. This is a fundamental step in determining the validity of the construct (Bond and Fox 2007). The PCAR analysis results show that the raw variance explained by measures of each sub-construct closely matched the expected target. The noise level of each sub-construct was acceptable because it was far from maximum of 15 % as recommended by Linacre (2003) and Fisher (2007).

Findings and Discussions

In total, 2304 questionnaires were returned. 121 of them were incomplete and questionnaires with double responses were dropped. Hence, the total of respondents became 2183. The female respondents were 19.6 % more than the male respondents. These constitute a total of 877 male students and 1306 female students. Table 1 is a leadership competency profile of students by gender, which shows the mean percentage of female respondents had higher scores for all sub-constructs within the competence of the leadership than male respondents.

An analysis was carried out to study the existence of GDIF. By using Winsteps, two-tailed t-test was performed to analyse GDIF. At the 95 % confidence level, the critical *t* value used to determine the level of significance was 2.0. In this study, GDIF analysis show statistical significant differences between boys and girls based on the levels of difficulty of items (Table 2). A negative GDIF index indicates that an

Table 1 Leadership competency profile by gender

Gender		Personality	Values	Competencies
Male	Mean	54.8425	58.8277	53.3145
	N	877	877	877
	S.D	6.71759	5.23824	4.97018
Female	Mean	54.9530	59.9988	53.8717
	N	1306	1306	1306
	S.D	5.82643	5.16579	4.75527

Table 2 Significant GDIF based on gender in personality

Person class	DIF measure	Person class	DIF measure	DIF contrast	*t*	Name of item
1	0.33	2	0.46	−0.13	−2.49	CP1
1	0.19	2	0.37	−0.18	−3.49	CP2
1	−0.31	2	−0.46	0.15	2.68	CP4
1	0.50	2	0.36	0.14	2.71	CP7
1	0.53	2	0.70	−0.17	−3.45	CP9
1	0.12	2	0.26	−0.14	−2.75	CP11
1	−0.44	2	−0.71	0.27	4.82	CP14

item is easier to be agreed by the males and vice versa. GDIF directions can be seen in graphs (Figs. 1, 2 and 3) where the horizontal axis shows each item in the constructs studied, while vertical axis shows the level of difficulty item (DIF measure) for males and females. Item difficulty level of male students is represented by

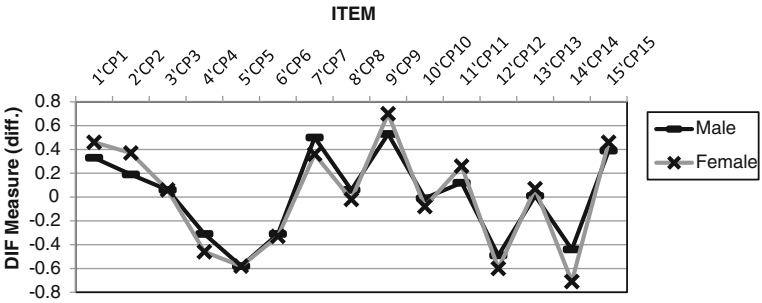


Fig. 1 GDIF plot of M3SLI (Personality)

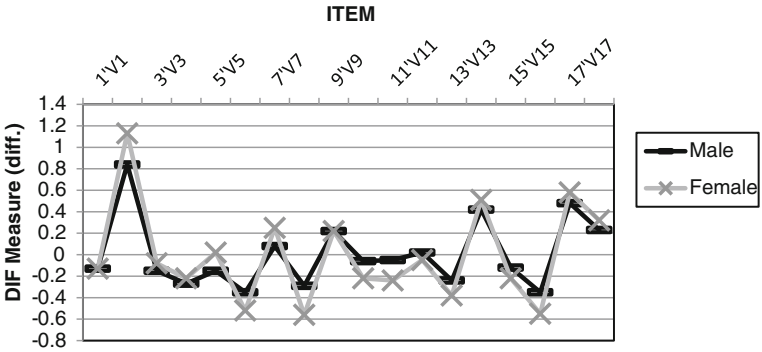


Fig. 2 GDIF plot of M3SLI (Values)

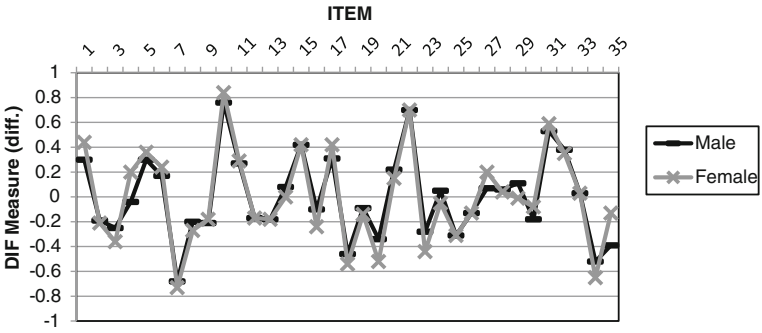


Fig. 3 GDIF plot of M3SLI (Competencies)

the symbol, while the girls are represented by the symbol **x**. Items that have a low level of difficulty in a group shows that the group is easier to agree with the item.

Figure 1 shows the DIF plot based on DIF measure for personality. Table 2 shows the results of GDIF analysis based on items in Personality construct. The analysed data demonstrates that seven items (47.7 %) from the 15 items in the personality construct show the significance of the GDIF in value $t \geq 2.0$ logit. The items are CP1, CP2, CP4, CP7, CP9, CP11 and CP14. The GDIF contrast shows that the seven items do not show serious GDIF because less than 0.5 logit. When DIF size not more than 0.5, it can be neglected (Bond and Fox 2007).

Table 3 shows the results of GDIF analysis based on items in Values construct. The analysed data demonstrates that 13 items (72.2 %) from the 18 items in the values construct show the significance of the GDIF in value $t \geq 2.0$ logit. The GDIF Contrast indicates that all the items do not show serious GDIF because is less than 0.5 logit. The items are V2, V5, V6, V7, V8, V10, V11, V13, V14, V15, V16, V17 and V18. Figure 2 shows the DIF plot based on DIF measure in Values.

Table 4 shows the results of GDIF analysis based on items in competencies construct. The analysed data shows that 11 items (31.4 %) from the 35 items show the significance of GDIF in value $t \geq 2.0$ logit. The GDIF Contrast (≥0.5 logit) shows that the 11 items do not show serious GDIF because the GDIF index is less than 0.5 logit. The items are LS1, LS3, LS4, LS16, LS17, LD20, LS23, LS27, LS29, LS34 and LS35. Figure 3 shows the DIF plot based on DIF measure in competencies.

As a summary, the result in Table 5 shows that out of the 31 items indicating the existence of LDIF, 14 items (45.2 %) were easier for urban area students, while 17 items (54.8 %) were easier for those from rural schools. There was a relatively large percentage in favour of both school locations in the measurement.

Table 3 Significant GDIF based on gender in values

Person class	DIF measure	Person class	DIF measure	DIF contrast	*t*	Name of item
1	0.84	2	1.13	−0.28	−6.85	V2
1	0.15	2	0.02	−0.17	−3.84	V5
1	−0.35	2	−0.52	0.17	3.45	V6
1	0.08	2	0.25	−0.17	−4.07	V7
1	−0.29	2	−0.56	0.27	5.46	V8
1	−0.06	2	−0.22	0.16	3.52	V10
1	−0.05	2	−0.24	0.18	4.09	V11
1	−0.24	2	−0.38	0.14	2.96	V13
1	0.42	2	0.51	−0.09	−2.12	V14
1	−0.12	2	−0.22	0.09	2.04	V15
1	−0.35	2	−0.55	0.20	4.05	V16
1	0.48	2	0.58	−0.10	−2.59	V17
1	0.23	2	0.32	−0.09	−2.06	V18

Table 4 Significant GDIF based on gender in competencies

Person class	DIF measure	Person class	DIF measure	DIF contrast	*t*	Name of item
1	0.30	2	0.44	−0.14	−2.81	LS1
1	−0.25	2	−0.36	0.11	2.18	LS3
1	−0.04	2	0.20	−0.24	−4.85	LS4
1	−0.10	2	−0.24	0.14	2.74	LS16
1	0.31	2	0.42	−0.11	−2.27	LS17
1	−0.34	2	−0.52	0.17	3.32	LS20
1	−0.28	2	−0.44	0.16	3.14	LS23
1	0.07	2	0.20	−0.13	−2.61	LS27
1	0.11	2	−0.01	0.12	2.47	LS29
1	−0.52	2	−0.65	0.13	2.46	LS34
1	−0.39	2	−0.13	−0.25	−4.98	LS35

Table 5 Direction of GDIF

Construct	Item administrated	Item with $t \geq 2.0$ logit	Direction of LDIF	
			Male	Female
Personality	15	7	3	4
Values	18	13	6	7
Competencies	35	11	5	6
Total	68	31	14	17

Discussion

The study shows that there are significant differences between both genders in responding to the sub-constructs (core personality, values and personality) as the *t*-values were at ±2.0. In core personality, male respondents agree easily to CP1 (I can handle stress well) and CP11 (I am emotionally stable, not easily influenced by surrounding.). Male respondents feel stronger stress arising from family factor compared to female respondents (Kai-wen 2009). The result of this study is contradicted with the study of Jaafar and Hidayah (2013) which showed that female respondents were able to deal with stress more effectively than male. They claimed that the male respondents were more affected by stress and male students by far lose their patience easily when they faced problems. Besides, the male respondents were found to display more emotional inhibition and low emotional stability than their female counterparts (Budaev 1999; Matud 2004).

Besides, Kwon and Song (2011) found that extraversion was a male-specific trait while openness to experience was a female-specific trait. In this study, extraversion was a trait that was dominated by female students because female appear to agree more with item CP4 (I am a talkative person), CP7 (I am a reflective person), CP9 (I

am a sociable person.) and CP14 (I am considerate to others.) if compared to the male respondents. On the other hand, openness to experience was dominated by male students because male respondents replied positively when responding to CP2 (I always have new ideas) in this study.

Prochaska et al. (1992 cited in Quinn and Spreitzer 2006) state that value leads to a sense of personal growth and awareness that an individual is becoming the kind of person s/he wants to be in the context of leadership. Hofmann (2009) agrees that value systems are directly related to individuals' world views which primarily hold conscious beliefs about how things are or should be. There were six items in values construct that indicate advantage to male respondents. The items are V2 (Wealth), V5 (Friendship), V7 (Freedom), V14 (Power), V17 (Influence) and V18 (Status). This indicates that male respondents believe that wealth, friendship, freedom, power, influence and status are more important values in performing leadership tasks in comparison to female respondents.

Men tend to place themselves in the hierarchy and authority for the sake of rank and status while women tend to form groups and collect power as a support network (Fisher 2000). Biologically, this can be due to the testosterone factor in males which cause them to fight for the rank and status. They usually believe that influence is important to ask other people to move. In performing leadership tasks, male respondents also believe that friendship and wealth are important too because leaders are not able to work alone and organisations that they lead need financial input to run their activities. In schools, some activities in the organisations cannot be organised because of lack of financial resources. Male respondents are also more concerned about wealth. Different from male, females are more anxious of taking risks when it comes to money and prefer to stay loyal to their existing providers (Pine 2009).

On the other hand, female were more agree with seven items respondents: V6 (Independence), V8 (Justice), V10 (Self-Direction), V11 (Obedience), V13 (Family), V15 (Truth) and V16 (Protection). Men's sense of self-worth is derived from providing for and protecting their families (Birkby and Harmon 2002). Therefore, in performing leadership tasks, male respondents are more concerned about protection to family while female seek for self-protection. Biologically, the oestrogen factor in female produce nurturing and connecting behaviours. Thus, female believes that family is an important values in performing leadership tasks. Furthermore, they believe that obedience, truth and justice are important in performing leadership tasks to gain respects from their members. Although Jepsen and Rodwell (2012) support this claim and assert that justice has a diffuse effect for males, but not for females, in this study, justice affects females more than males.

Although in most parts of the country, female leaders are still far behind, there are more female leaders than ever before in the top leadership positions (Latu et al. 2013). Significantly, more females obtain supervision and middle management positions, but having said that becoming elite leaders or senior management is still relatively rare (Eagly and Karau 2002). In most setting, females possess lower level of status and power than male. This leads to the perception caused people to assume that males are more competent and knowledgeable than females (Carli 2001). Currently, it can be seen that female fight for and recognition for their roles in

society, including leadership. Besides, they also fight for recognition to prove that they can lead effectively. This indirectly inspires female students in high school to involve in leadership. Therefore, female respondents in this study easier to agree that independence and self-direction are important values in performing leadership tasks.

In competencies construct, there were six items which showed advantage to female respondents and five items to male. Within the six items, male respondents easier to agree that they are alert to the change of the global need (LS1), could handle their emotion well (LS4), show rather than tell the ways (LS17), could cope well with changes (LS27), and could solve problem well (LS35). Results of this study correspond to the study of Prime et al. (2009). They claims that male were more inclined towards action oriented, 'take-charge' leadership behaviours. Therefore, they suggested that male can solve problems better than females.

Prime et al. (2009) found that females were more effective than males at care-taking leadership behaviours. In line with that study, females in this study were easier to agree that they could lead according to situation (LS3), accept different types of idea from members (LS16), ready to change with the support of the members (LS20), could overcome obstacles patiently (LS23), could plan activities of my club/class together with the members (LS29) and ready to change with the support from my teachers (LS34). In other words, female respondents agreed that they could communicate better and leading changes with members in their organisations.

Conclusion

As a summary, the items examined in three constructs tended to be bias towards female respondents. Overall, M3SLI shows 17 items bias towards females while 14 items bias towards males. Nonetheless, the contrasts between the groups were not serious as the GDIF index showed less than 0.5 logit. Therefore, all the items were maintained. DIF analysis is important to ensure that the instrument is valid and less biased to certain group of respondents. Through DIF, items with extreme levels of DIF were identified and omitted. Further studies are needed to understand the difference in DIF items according to the respondents' grades, leadership posts and school locations.

References

Amirianzadeh, M., Jaafari, P., Ghourchian, N., & Jowkar, B. (2010). College student leadership competencies development: A model. *International Journal for Cross-Disciplinary Subjects in Education, 1*(3), 168–172.

Birkby, S. J., & Harmon, S. P. (2002). Special report. *Radiographics, 22*(4), 907–909.

Bisland, A. (2004). Developing leadership skills in young gifted students. *Gifted Child Today*, *27* (1), 24–27. Retrieved from http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ682651.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, New Jersey: Routledge.

Budaev, S. V. (1999). Sex differences in the big five personality factors: Testing an evolutionary hypothesis. *Personality and Individual Differences, 26*, 801–813.

Carli, L. L. (2001). Gender and social influence. *Journal of Social Issues*, *57*(4), 725–741. http://doi.org/10.1111/0022-4537.00238.

Connerley, M. L., Mecham, R. L., & Strauss, J. P. (2008). Gender differences in leadership competencies, expatriate readiness, and performance. *Gender in Management: An International Journal*, *23*(5), 300–316. http://doi.org/10.1108/17542410810887347.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. http://doi.org/10.1037//0033-295X.109.3.573.

Farver, C. S. (2007). *Student leadership self-perception compared to teacher perception of student leaders, achievement scores, and student demographics in a cross-section of students in third, sixth, and ninth grades*. Drake University.

Fertman, C. I., & Van Linden, J. A. (1999). Character education for developing youth leadership. *National Association of Secondary School Principals Bulletin, 83*(605), 11–16.

Fisher, H. (2000). The natural leadership talents of women. In L. Coughlin, E. Wingard, & K. Hollihan (Eds.), *The first sex: The natural talents of women and how they are changing the world* (pp. 133–139). Ballantine Books.

Fisher, J. W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*, 1095.

Haslina, H. (2013). *Pembinaan Instrumen Kesediaan Belajar Awal Kanak-kanak (IKBAK)*. Universiti Sains Malaysia.

Hassan, J., & Safar, A. S. (2010). *Pembinaan Kecemerlangan Diri Pimpinan Pelajar Menerusi Penglibatan Dalam Aktiviti Kokurikulum Di Universiti Teknologi Malaysia, Skudai.* Skudai.

Hine, G. (2011). *Exploring the development of student leadership potential within a catholic school: A qualitative case study*. University of Notre Dame Australia. Retrieved from http://researchonline.nd.edu.au/cgi/viewcontent.cgi?article=1062&context=theses.

Hofmann, J. G. (2009). *The multidimensional structure and function of human values*. University of Southern California.

Ismail, K. H., Anwar, K., Ahmad, S., Selamat, J. H., & Ahmad, A. (2013). Personality profile of students' council: A comparative study between genders. *Asian Social Science, 9*(4), 77–84. doi:10.5539/ass.v9n4p77

Jaafar, M., & Hidayah, N. (2013). Exploring stress with focus on students in a higher educational institution: A case study in universiti. *International Surveying Research Journal, 3*(1), 17–31.

Jepsen, D. M., & Rodwell, J. (2012). Female perceptions of organizational justice. *Gender, Work & Organization*, *19*(6), 723–740. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0432.2010.00538.x/abstract.

Johnson, S. K., Murphy, S. E., Zewdie, S., & Reichard, R. J. (2008). The strong, sensitive type: Effects of gender stereotypes and leadership prototypes on the evaluation of male and female leaders. *Organizational Behavior and Human Decision Processes*, *106*(1), 39–60. http://doi.org/10.1016/j.obhdp.2007.12.002.

Kai-wen, C. (2009). A study of stress sources among college students in Taiwan. *Journal of Academic and Business Ethics, 2*, 1–8.

Kwon, N., & Song, H. (2011). Personality traits, gender, and information competency among college students. *Malaysian Journal of Library & Information Science, 16*(1), 87–107.

Latu, I. M., Schmid, M., Lammers, J., & Bombari, D. (2013). Successful female leaders empower women's behavior in leadership tasks. *Journal of Experimental Social Psychology, 49*, 444–448.

Linacre, J. M. (2003). *Dimensionality: contrasts & variances*. Retrieved January 20, 2015, from http://www.winsteps.com/winman/principalcomponents.htm.

Linacre, J. M., & Wright, B. D. (2012). *A user's guide to WINSTEPS ministeps Rasch model computer programs*. Chicago: Mesa Press.

Matud, M. P. (2004). Gender differences in stress and coping styles. *Personality and Individual Differences*, *37*(7), 1401–1415. http://doi.org/10.1016/j.paid.2004.01.010.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *The American Psychologist*, *28*(1), 1–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4684069.

Minckler, C. H. (2011). *Teacher social capital : The development of a conceptual model and measurement framework with application to educational leadership and teacher efficacy*. University of Louisiana.

Ministry of Education Malaysia. (2012). *Malaysia education blueprint 2013–2025*. Kuala Lumpur: Kementerian Pelajaran Sabah.

Mustamin, M. A. M. H. M. Y. (2013). Perbandingan kompetensi pengetua sekolah antara Malaysia dan Indonesia. *Jurnal Teknologi (Sciences and Engineering), 62*(1), 7–16.

Othman, N. B., Salleh, S. B., Hussein, H., & Wahid, H. B. A. (2014). Assessing construct validity and reliability of competitiveness scale using Rasch model approach. In *Proceedings of the 2014 WEI International Academic Conference* (pp. 113–120). Bali, Indonesia.

Paustian-underdahl, S. C., Walker, L. S., & Woehr, D. J. (2014). Gender and perceptions of leadership effectiveness: A meta-analysis of contextual moderators. *Journal of Applied Psychology, 99*(6), 1129–1145.

Pine, K. J. (2009). *Report on a survey into female economic behaviour and the emotion regulatory role of spending*. Hatfield, UK.

Posner, B. Z. (2012). Effectively measuring student leadership. *Administrative Sciences*, *2*(4), 221–234. http://doi.org/10.3390/admsci2040221.

Prime, J. L., Carter, N. M., & Welbourne, T. M. (2009). Women "Take Care," men "Take Charge": Managers' stereotypic perceptions of women and men leaders. *The Psychologist-Manager Journal*, *12*, 25–49. http://doi.org/10.1080/10887150802371799.

Prochaska, J. O., DiClemente, C. C., & Norcross, J. C. (1992). In search of how people change: applications to addictive behaviors. *American Psychologist, 47*(9), 1102–1114. doi:10.3109/10884609309149692

Quinn, R., & Spreitzer, G. (2006). Entering the fundamental state of leadership: A framework for the positive transformation of self and others. In R. Burk & C. Cooper (Eds.), *Inspiring leaders* (pp. 9–10). London: Routledge.

Rahayah, A. S., Ibrahim, S. I., Malek, N. H., Sarif, S. H., & Yassin, S. F. M. (2010). Gender differential item functioning (GDIF) in an online intelligencetest. In *WSEAS International Conference on Education and Educational Technology (EDU'10)* (pp. 331–335).

Renn, K. A., & Ozaki, C. C. (2010). Psychosocial and leadership identities among leaders of identity-based campus organizations. *Journal of Diversity in Higher Education*, *3*(1), 14–26. http://doi.org/10.1037/a0018564.

Rosseni, D., Mazalah, A., Faisal, M. K., Norhaslinda Mohamad, S., Aidah Abdul, K., Nur Ayu, J., et al. (2009). Kesahan dan kebolehpercayaan soal selidik gaya e-pembelajaran (else) versi 8.1 menggunakan Model Pengukuran Rasch. *Jurnal Pengukuran Kualiti Dan Analisis*, *5*(2), 15–27.

Tubbs, S., & Schulz, E. (2006). Exploring a taxonomy of global leadership competencies and meta-competencies. *Journal of American Academy of Business, 8*(2), 29–35.

Zakaria, S., Aziz, A. A., Mohamed, A., Arshad, N, H., Ghulman, H. A., & Masodi, M. S. (2008). Assessment of information managers' competency using Rasch measurement. *Proceedings—3rd International Conference on Convergence and Hybrid Information Technology, ICCIT 2008*, *1*, 190–196. http://doi.org/10.1109/ICCIT.2008.387.

Zamru, M., & Anisah, A. (2012). Kesahan dan kebolehpercayaan mygsi dalam mengukur kemahiran komunikasi, kepimpinan dan kerja berkumpulan guru Pelatih Bahasa Melayu (The validity and reliability of mygsi in measuring communication skills, leadership and teamwork among Malay language T). *Jurnal Pendidikan Malaysia, 36*(1), 67–75.

Measurement as a Medium for Communication and Social Action I: A Phenomenological View of Science and Society

William P. Fisher Jr. and Robert F. Cavanagh

The Argument for a Phenomenological Approach

The point of entrée into this work is a view that science and scientific progress are inextricably linked with the role of technology in history and societal development. Sociological accounts of science have been transformed in recent studies of the history and philosophy of science, away from references to the social as causative to a new sense of the social as the product of technologically embodied and locally situated relationships (Crease 2011; Latour 1987, 1990, 2005, 2013). For example, the Strong Programme of the 1970s moved to replace the philosophy of science with a sociology of science (Bloor 1976). The resulting relativism led to a frustrating sense that "anything goes," with renewed efforts to either dig deeper into existing philosophies or to move past the modern-postmodern, and positivist-anti-positivist divides. Accordingly, in many quarters, positivist conceptions of science that were developed centuries earlier in the Enlightenment persist in "popularist" notions of science, such as naïve realism (Michell 2003).

In the same way that some choose deeper entrenchment in positivism, Kampen and Tobi (2011), like many others before them (for instance, Martin and Sugarman 2001; Bryman 2007), are willing to allow the natural sciences a positivist cast. They espouse an irreconcilable ontological difference between the human and natural sciences, commenting that, "…there is a sharp divide between interpretavist and neo-positivist ontological world views which cannot be expected to be resolved in the near future" (Kampen and Tobi 2011, p. 1). This plurality has implications for

W.P. Fisher Jr. (✉)
University of California, Berkeley, CA, USA
e-mail: wfisher@berkeley.edu

R.F. Cavanagh
Curtin University, Bentley, WA, Australia
e-mail: R.Cavanagh@curtin.edu.au

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_10

the function of measurement in science. Instrumentalist or deterministic measurement models are commonly associated with positivist ontologies; and probabilistic measurement models can be more productively associated with anti-positivist or post-positivist ontologies or stochastic world views. The strength of association depends on the extent of congruence between respective ontological assumptions and the theory within particular measurement models. Similarly, the ontologies informing human science, social science and humanistic inquiry are related to the respective theoretical bases for measurement in these three fields. So shifts in philosophical orientation accompanying movement between different forms of enquiry require concomitant reframing of the existing theory of measurement or possibly development of a new theory. Adoption of a philosophical perspective on measurement theory has the potential to provide new insights of theoretical significance to measurement. This project adopts a philosophical view of measurement and uses a phenomenological lens.

There are several reasons for the choice of phenomenology as the frame of reference for working out ways past, over, or through the ontological divide between positivism and anti-positivism. Some of these stem from a need to illuminate the three major substantive concerns of the project; concerns about measurement, meaningful communication, and societal renewal. More specifically, phenomenology is implicated in

- proposing hermeneutical insight into the reading of scientific measuring instruments (Heelan 1983);
- emphasizing meaning in communication, and the "commonality of language ensuring a shared acceptance of meaning and ability to vocalise thoughts" (Regan 2012, p. 288); and
- recognition of lifeworld as a fundamental construct in society, "an historically and culturally invariant structure, without which human life and its various modes of experience would be unimaginable" (Schieman 2014, p. 32).

Another reason for choosing phenomenology is its basis in geometry as a root model of scientific conduct (Husserl 1954/1970; Gadamer 1980, pp. 100–101). A major intention of Husserl's sense of pure or transcendental phenomenology is the categorisation of lived experiences and mental activities in order to develop an understanding of underlying order or coherence (Husserl 1913/1983). This process is analogous to the development of the schema constituting the natural sciences, for example, the taxonomies of Biology and Geology. Reduction into linguistic expressions distils defining features proven invariant within the limits of the stochastic nature of the subjective experiences. Yet other reasons stem from the need to strengthen the philosophical foundations of the study of measurement, particularly those underpinning the discipline of metrology as it is emerging in the social and human sciences (Fisher and Stenner 2011; Fisher and Wilson 2015; Maul et al. 2016).

This first paper, *A Phenomenological View of Science and Society*, elaborates on these rationales. A multidimensional analytic frame was applied to inform assaying of classical and modern phenomenology. The frame comprises four characteristics

of phenomenology: *Back to the Things Themselves*; *Authentic Method*; *Unity of Subject and Object*; and *The World of the Text*. The essence of each was distilled to identify the qualities considered essential when specifying a science grounded in phenomenological precepts. The paper concludes with the introduction of the notion of an unmodern or amodern perspective on measurement.

First Characteristic of Phenomenology: Back to the Things Themselves

Phenomenology describes phenomena as they are manifest in the convergence of things themselves with the consciousness of the experiencer. It requires the phenomena to be understood in their own terms, which in the sense of phenomenology advanced by Husserl and retained in the Husserlian tradition, focuses on a transcendental consciousness, a mode of thought not influenced by preconceptions, misconstructions, or the imposition of explanations, including scientific theories (Moran 2000). Heidegger (1962, p. 50), however, took a different path, and elaborates:

> Thus the term 'phenomenology' expresses a maxim which can be formulated as 'To the things themselves!' It is opposed to all free-floating constructions and accidental findings; it is opposed to taking over any conceptions which only seem to have been demonstrated; it is opposed to those pseudo-questions which parade themselves as 'problems', often for generations at a time.

In Heidegger's view, this opposition does not necessitate the rejection or dismissal of fore-structures. Instead, they are understood by reference to the experience:

> Our first, last, and constant task is never to allow our fore-having, fore-sight, and fore-conception to be presented to us by fancies and popular conceptions, but rather to make the scientific theme secure by working out these fore-structures in terms of the things themselves (Heidegger 1962, p. 195).

Husserl (1911/1965, p. 108) emphasizes the absolute nature of such a conception in contrast to conceptions derived from relativistic comparisons and contrasts:

> 'To the things themselves!' Not, however, to things as they are 'in themselves' (*an sich*), where their being is relative, but in the psychic flow, where their being is absolute, an essential being with the absoluteness of subjectivity.

Heidegger, on a page in his personal copy of a book of Husserl's emphasizing the return to the things themselves, in frustration with what he saw as the inconsistency of advising a return to the things themselves with Husserl's overriding focus on transcendental consciousness, wrote in the margins, "Let us take Husserl at his word!" (Gadamer 1991a, p. 14). Phenomenological reduction is employed to identify features of our experience that are both necessary and invariant, the eidos or essence of a particular subjective experience. But where Husserl conceives

reduction in terms of a bracketing of presuppositions, as though all of them can be identified and set aside through an effort of will, Heidegger conceives reduction in terms of the way things come into words. A potentially infinite array of variations in experience is expressed in a spoken or written form of limited length. This first moment in the phenomenological method is then applied in the construction of meaning, the second moment, which in turn may eventually be followed by a dismantling or deconstruction of the original concept, with the aim of resolving inconsistencies and returning to a new reduction (Heidegger 1982, pp. 19–23, 320–330; Fisher and Stenner 2011).

The phenomenological method is, then, complemented by the way prejudgments and moving closer to the "thing itself" are experienced in the hermeneutic circle:

> Each circle - or cycle - follows a sequence of four phases - a. *experiencing/observing,* b. *theory-making*, c. *theory-testing,* and d. *deciding* - each phase giving access to new insights; each cycle leading to a partially transformed beginning of a new cycle in which further development is made. Each cycle revises and improves the previous cycles of inquiry until the basic queries have been sufficiently explored dialogically. (Heelan 2014, p. 95)

In summary, the notion of "back to the things themselves" has implications for how we approach the object of inquiry: "we always come to our object of study with a set of prejudgments: an idea of what the problem is, what type of information we are looking for, and what will count as an answer" (Frodeman 2014, p. 74). It also has implications for how we exploit emergent understanding of the object of inquiry: "we remain open to correction, allowing the text or object to instruct us and suggest new meanings and approaches" (Frodeman 2014, p. 74).

Second Characteristic of Phenomenology: Authentic Method

Genuine method is intimately integrated with the way we come to terms with things themselves and with the principles on which science is conducted. Working through the presuppositions brought to bear in observation in terms of the things themselves means subjecting thought to the activity, behavior or movement of the object of interest. Instead of applying subjective, externally determined processes to a separate object, authentic method begins from a unity of subject and object caught up together in a flow of experience (Gadamer 1989, pp. 463–464). Method as a concept is a following along after (*meta-*) on the path (*odos*) of the thing itself experienced in thought via interactions with it (Heidegger 1991, p. 63; Gadamer 1989, pp. 459–461).

Phenomenology is primarily a way of conceptualizing method in this sense (Heidegger 1962, p. 50, 1977, p. 32), where we come into contact with things before they have been fixed as abstract, theoretical, conceptual entities removed from the concrete local experience of human praxis, history, and culture (Heelan 1994, p. 369;

Gadamer 1976). All observations are informed by ideas that focus attention and that effectively model things in the world in particular ways. Scientific observations are informed by models positing what are in fact unrealistic ideals (Butterfield 1957, pp. 16–17), such as Galileo's sense of the uniform acceleration of perfectly smooth balls rolling on frictionless planes, or Rasch's sense of reading comprehension being a function only of the reader's ability and text complexity. To communicate and work with unrealistically ideal models, they have to be embodied in linguistic and technological forms. Western philosophy originated in the capacity to look through geometric and numeric figures illustrating and representing mathematical relationships to those relationships themselves (Gadamer 1980, pp. 100–101; Heidegger 1977). Much remains to be done in the way of extending these philosophical considerations into the domains of mathematical psychology and sociology (Fisher 1992, 2003, 2004).

Newton's laws, and the resulting textbook engineering methods, have succeeded in uprooting conceptualizations of mass, force, and acceleration from the undifferentiated earth of "the concrete plurality of particular existents" (Gadamer 1976, p. 9) that remained systematically ungrasped and unrepresented in premodern history. Heidegger (1977, p. 32) uses poetic language to describe "catching sight of what comes into presence in technology," and asking "how the instrumental comes to presence as a kind of causality," points us toward possibilities for experiencing "this coming to presence as the destining of a revealing." In other words, when we pay close attention to the ways in which technologies (overtly scientific instruments, as well as alphabets, grammars, and syntax of language) frame experience and perception, we apprehend something methodologically important, the projecting into history of something new coming into words.

But authentic method is not a panacea. Both the strength and the danger of method lies in the way it narrows thought and focuses attention:

> There always remains the constant danger of the systematic problem of philosophy itself: that the part of lived reality that can enter into the concept is always a flattened version-like every projection of a living bodily existence onto a surface. The gain in unambiguous comprehensibility and repeatable certainty is matched by a loss in stimulating multiplicity of meaning. (Gadamer 1991b, p. 7)

In other words, "all interpretation makes its object univocal and, by providing access to it, necessarily also obstructs access to it" (Gadamer 1991b, p. 8; also see Gasché 2014). To the extent a method is authentic, when we succeed in abstracting general concepts from the mixtures of undifferentiated experiences, we also tend to selectively ignore everything that does not fit.

This unidimensional leveling of differences is not, however, completely inevitable, necessary, or total. If it were, we would have no concept of exceptions that prove (in the sense of test) the rule, or of the anomalies that accumulate and call for the reconceptualization of one or another domain of experience. In quite fundamental ways, stochastic processes signal the demise of reductionism, as they are unavoidable even in areas as seemingly deterministic as arithmetic and Newtonian

physics (Chaitin 1994), and the implementation of uniform laboratory standards (Berg and Timmermans 2000). We will return to these issues of authentic method, especially in the context of the world of the text.

Third Characteristic of Phenomenology: Unity of Subject and Object

The unity of subject and object characterizing phenomenology rejects separation between mind and world, between language and reality, and between subject and object. In order to understand the ontological and epistemological consequences of rejecting the dichotomies, it is informative to examine their origins. The history of the separation extends back to ancient Greece and can be tracked through the enlightenment into the modern era. The Greek philosopher Plato advocated the need to transcend human knowledge and for our minds to represent a reality that exists independently of our minds. This is a metaphysical realist view of the natural world. While "desires" and "reason" existed in the Greek culture, as Hegel (1910/2003) noted, these were in harmony. Hegel also noted the harmony persisted until the emergence of "individual conscience" in protestant Europe in the eighteenth century and the rise of the "new science." The new science was in tension with the "omniscience of the metaphysical tradition," there was an imbalance between "a science of reason based on concepts and a science based on experience" (Gadamer 1970/2006, p. 16). A solution to this problem was found by Kant who formulated another dualism:

> By getting straight the distinction between sensibility and the understanding and by keeping straight the sources of our concepts, we can protect the claims of geometry from those of metaphysics. Geometry applies to the objects of sensibility, objects given in space and time, metaphysics applies to the objects of the understanding, that is, God and moral perfection. (Carson 2011, p. 30)

Kant also contrasted metaphysical idealism with realism: "Metaphysical idealism is the view that the ultimate nature of reality is constituted by minds or ideas. Realism holds, on the contrary, that the nature of reality is mind-independent" (Dudley and Engelhard 2011, p. 3). A related more general issue is the relation between one's experiences, and the world, that is, between phenomena and reality. Dilworth (2007, p. 9) attested to the attention this has been given:

> At one time or another virtually every conceivable line has been taken on the issue, from the view that there is no reality other than phenomena [empiricism], to the view that reality, while different from phenomena, alone causes and is perfectly represented by them [realism].

For example, empiricism is "broadly speaking, the view that scientific investigation be confined to phenomena and their formal relations" (Dilworth 2007, p. 9). An extreme form of empiricism is the positivism developed by Comte. He was insistent that understanding nature and discovering its laws must commence with,

and be restricted to, the analysis of phenomena. Ontological questions are neither asked nor answered; instead the focus is epistemological concerns, particularly how to develop theory from observations. A central tenet of logical positivism/empiricism is the theory/observation distinction. "It is only because observations are independent of theories that they could serve as evidential warrants to appraise the adequacy of theories, to ground theory comparisons" (Zammito 2004, p. 10). The fidelity of the distinction and its empirical consequences are threatened by the more recent recognition that observations cannot be completely independent of theory, the theory-laden nature of observations (Shapere 1984). Another criticism of positivism is its inadequacy in understanding human behavior, particularly when applied in the social sciences.

> Positivist social science is an impossible construction for human inquiry. Not only does it belie a bureaucratic market mentality (research is big business), but in its legitimation of social structures and practices that deny intersubjective meanings, it fails as a discourse for personal agency, moral obligation and political responsibility. (Brieschke 1992, p. 178)

Kuhn (1970) recognized the limitations of the assumption of universality underpinning positivism.

> Kuhn noted that scientific methods and practices were not universal, but localized within quite tightly bounded communities of practitioners. He acknowledged that phenomena could not be observed raw, but were always interpreted through a framework of preconceptions and according to assumptions bound up with the use of certain instruments. And he recognized that, while paradigms guided scientific research, they did not determine what sense could be made of new experiences. (Golinski 2012, p. 31)

These criticisms of positivism constitute some of the arguments of anti-positivism that led to a move beyond positivism to post-positivism, sometimes called post-modernism. However, the boundaries between positivism, anti-positivism and post-positivism are difficult to define. Cohen (1989 cited in Heidtman et al. 2000, p. 2) illustrates the confusion:

> If Positivism means a commitment to using evidence, then this author is a Positivist; if it means that nonobservable entities are inadmissible, then the present writer is an Anti-Positivist. If Post-Positivism represents a concern with the theoretical relevance of observables, then this analyst is a Post-Positivist; and so on.

Notwithstanding, Heidtman et al. (2000, p. 17) identified three principles of a post-positivist perspective: "All scientific data are theoretically informed," "empirical commitments are not based solely on experimental evidence"; and "fundamental shifts of scientific belief occur only when empirical changes are matched by the availability of alternative theoretical commitments." The principles posit that science does not proceed through inductive processes or that a theory can ever be conclusively validated by empirical means (see Kuhn 1970; Lakatos 1970). This is a unified orientation because in "rejecting the epistemological distinction between observation statements, grounded in experience, and theoretical statements, based on conjecture, post-positivists identified knowledge with theory" (McEvoy 2007, p. 386).

What remains missing in this anti-positivist perspective is the role of technologically embodied knowledge (Dewey 2012; Galison 1997; Golinksi 2012; Heelan 1983; Ihde 1991; Latour 1990, 1993, 2013). Technologies and instruments, including phonemes, movable type, diagrams, books, and thermometers, must be accounted for, since they embody the media through which meanings are communicated and shared.

The dualistic reasoning of classical Western philosophy (modernism) and the fragmentation of separated subjects and objects are shown to be fundamentally flawed by deconstructions elaborating the unified subject–object approaches typical of phenomenology [anti-positivism, in Galison's (1997) terms]. Gadamer (1989, p. 459), for instance, points out that,

> In this thinking [of Plato and Aristotle presuming method as dialectically absorbing thought into the movement of things themselves] there is no question of a self-conscious spirit without world which would have to find its way to worldly being; both belong originally to each other. The relationship is primary.

Accordingly, "Dialectic, this expression of the logos, was not for the Greeks a movement performed by thought; what thought experiences is the movement of the thing itself" (Gadamer 1989, p. 460). And so, "We are simply following an internal necessity of the thing itself if we go beyond the idea of the object and the objectivity of understanding toward the idea that subject and object belong together" (Gadamer 1989, p. 461). What we are doing, then, is "…thinking out the consequences of language as medium" (Gadamer 1989, p. 461), moving past the modern and the postmodern to an unmodern (Dewey 2012) or amodern (Latour 1990, 1993, 2013) embodiment of understanding in the fused horizons of unified subject–objects.

Fourth Characteristic of Phenomenology: The World of the Text

When we intentionally focus on things themselves in relation to words and concepts, when method is understood as the activity of the thing itself experienced in thought, and when fused subject–object horizons are embodied in the technical media of words and instruments, we arrive at a productive perspective on Ricoeur's sense of the world of the text. As Ricoeur (1981, pp. 192–193) puts it, in appropriating meaning,

> …what is 'made our own' is not something mental, not the intention of another subject, nor some design supposedly hidden behind the text; rather, it is the projection of a world, the proposal of a mode of being in the world, which the text discloses in front of itself by means of its non-ostensive references. Far from saying that a subject, who already masters his own being-in-the-world, projects the a priori of his own understanding and interpolates this a priori in the text, I shall say that appropriation is the process by which the revelation of new modes of being—or if you prefer Wittgenstein to Heidegger, new 'forms of life'—gives the subject new capacities for knowing himself. If the reference of a text is the

> projection of a world, then it is not in the first instance the reader who projects himself. The reader is rather broadened in his capacity to project himself by receiving a new mode of being from the text itself.

Everyday language has the capacity to make ostensive references, pointing to features of the environment shared by speakers situated in a common location. But writing makes non-ostensive references, pointing at people, places, and things disconnected from the here and now of both the reader and the writer, introducing a new distance between signs and referents and broader horizons within which interlocuters belong to a shared community. Ricoeur describes this "distanciation phenomenologically, from the perspective of hermeneutics; we are more interested in it as a social process, to produce, through a sequence of mediations, the embedded system of meanings" (Taylor et al. 1996, p. 35). Latour (1987, p. 25) similarly takes up this distanciation as a key marker of factuality, pointing out that the more the credibility or meaning of a statement depends on who said it, and when and where, the less generalizable it is (Taylor et al. 1996, pp. 35–36).

Learning from a text involves the capacity to bring broad, contextual, linguistic expectations and specific, local expectations to bear in a way that both allows those expectations to be satisfied and makes them fluid and alive to new possibilities in the moment. The reader's lived world horizons are broadened via a pragmatic expansion of behavioral options opened up by the text. Different possible foundations for decisions, greater compassion for the plights of others, more forbearance in the face of complex circumstances, and innumerable other directions for action can follow from the reading of a text, whether it is a novel, a poem, a thermometer, a speedometer, or a look on someone's face.

The enframing of the world accomplished by language lifts the burden of initiation (Gadamer 1989, p. 104) from the reader by absorbing her or him into the play of signifiers, thereby providing new possibilities for thinking and acting. And it is here, where the advance work performed by language in making the world thinkable becomes apparent and useful, that the pragmatic overlap of the hermeneutic and the sociotechnical resides, with one caveat. For Ricoeur (1981, p. 191), "The ideality of the text remains the mediator in this process of the fusion of horizons," but for us, that ideality, like Latour's (1986, pp. 7–14) sense of the "immutable mobile," has to be recast as a boundary object seen in different ways from the varying perspectives of every stakeholder interacting with it (Star and Griesemer 1989; Fenwick 2010, p. 129; Fisher and Wilson 2015; Gooday 1997, p. 411). That said, Ricoeur (1981, p. 219), quite in harmony with Latour (2005), well states the fact that:

> …the function of substituting signs for things and of representing things by the means of signs, appears to be more than a mere effect in social life. It is its very foundation. We should have to say, according to this generalized function of the semiotic, not only that the symbolic function is social, but that social reality is fundamentally symbolic.

Like Gadamer's, Ricoeur's investigations stop with this realization of language as the medium of social life. Each understands in his own way that "the use and development of language is a process which has no single knowing and choosing

consciousness standing over against it" (Gadamer 1989, p. 463). But instead of taking up the question Hayek (1948, p. 54) regarded the central question of all social science, Gadamer and Ricoeur both choose to focus on what they consider the hermeneutic event proper, the coming into language of what has been said in the tradition, an event that is simultaneously appropriation and interpretation, the act of the thing itself that thought experiences.

Latour, however, goes in the opposite direction, implicitly taking up Hayek's (1948, p. 54) question, which Hayek posed, asking,

> How can the combination of fragments of knowledge existing in different minds bring about results which, if they were to be brought about deliberately, would require a knowledge on the part of the directing mind which no single person can possess? To show that in this sense the spontaneous actions of individuals will, under conditions which we can define, bring about a distribution of resources which can be understood as if it were made according to a single plan, although nobody has planned it, seems to me indeed an answer to the problem which has sometimes been metaphorically described as that of the 'social mind.' But we must not be surprised that such claims have usually been rejected, since we have not based them on the right grounds.

In their pursuit of answers to Hayek's question, Latour, Hutchins, and others working in science and technology studies have documented in exacting detail multiple instances of the processes through which metrological standards and traceability to them have brought about results via locally and spontaneously coordinated decisions and behaviors that nonetheless appear to follow a single centrally administered plan. Their pragmatic focus on what is said and done in the reading of instruments and the writing of memos, grant applications, conference presentations, reviews, letters of recommendation, and publications provides a wealth of material on the ways in which worlds are projected in front of texts, and are inhabited, even by those unversed in the language of mathematics that the Book of Nature is written in.

An Unmodern or Amodern Frame of Reference

Following Einstein's insight that major problems cannot be solved from within the frame of reference that provoked them, a profoundly different way of thinking and acting is required to initiate a new paradigm of scientific productivity, measurement, and innovation in the social sciences. The consequences of language as medium, as knowledge embodied in the technologies of standardized alphabets, grammars, phonemes, syntaxes, printing presses, books, web pages, and digital fonts, stands in radical contrast with the "fatal conceit" (Hayek 1988) of the modern Cartesian presumption of an independent subject making its own way to worldly being. Continued reliance on modern and postmodern conceptions advocating or criticizing subjectivities over against objects prevents us from formulating the concepts, methods, and tools needed for paradigm-shifting broad scale improvements in the quality of psychological and social measurement.

An alternative unmodern (Dewey 2012) or amodern (Latour 1990, 1993) frame of reference offers a fundamentally different basis for thinking about science and doing measurement. This alternative focuses

- on knowledge as technology,
- on the lack of a central authority over the use and development of language,
- on its recognition of end users as having little or no understanding of how language and technology work,
- on its acceptance of genuine method as a playful captivation in the flow of mutually implicated subjects and objects, and
- on its focus on the wide distribution of standardized tools as providing the language unifying fields of research and practice.

Thus, rather than continue waiting indefinitely for the modern project to arrive at its perpetually deferred fulfillment in an complete picture of the objective world, the unmodern perspective suggests we should instead define the terrain, the equipment, and the rules, roles, and responsibilities of teams and players in the language game of measurement. These matters will be further explored in the second paper, *The Promise and Power of Being Amodern.*

References

Berg, M., & Timmermans, S. (2000). Order and their others: On the constitution of universalities in medical work. *Configurations, 8*(1), 31–61.

Bloor, D. (1976). *Knowledge and social imagery*. London: Routledge and Kegan Paul.

Brieschke, P. A. (1992). Reparative praxis: Rethinking the catastrophe that is social science. *Theory into Practice, 31*(2), 173–180.

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research, 1*(1), 8–22.

Butterfield, H. (1957). *The origins of modern science*. New York: The Free Press.

Carson, E. (2011). Sensibility: Space and time, transcendental idealism. In W. Dudley & K. Engelhard (Eds.), *Immanuel Kant: Key concepts* (pp. 228–244). Durham: Acumen.

Chaitin, G. J. (1994). Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos, 4*(1), 3–15.

Crease, R. (2011). *World in the balance: The historic quest for an absolute system of measurement*. New York: W. W. Norton & Co.

Dewey, J. (2012). *Unmodern philosophy and modern philosophy.* (P. Deen, Ed.). Carbondale, Illinois: Southern Illinois University Press.

Dilworth, C. (2007). *The metaphysics of science: An account of modern science in terms of principles, laws and theories* (2nd ed.). Dordrecht, The Netherlands: Springer.

Dudley, W., & Engelhard, K. (2011). Introduction. In W. Dudley & K. Engelhard (Eds.), *Immanuel Kant: Key concepts* (pp. 1–10). Durham: Acumen.

Fenwick, T. J. (2010). ("Un")Doing standards in education with actor-network theory. *Journal of Education Policy, 25*(2), 117–133.

Fisher, W. P, Jr. (1992). Objectivity in measurement: A philosophical history of Rasch's separability theorem. In M. Wilson (Ed.), *Objective measurement: theory into practice* (Vol. I, pp. 29–58). Norwood, New Jersey: Ablex Publishing Corporation.

Fisher, W. P, Jr. (2003). Mathematics, measurement, metaphor, metaphysics: Parts I & II. *Theory & Psychology, 13*(6), 753–828.
Fisher, W. P, Jr. (2004). Meaning and method in the social sciences. *Human Studies: A Journal for Philosophy and the Social Sciences, 27*(4), 429–454.
Fisher, W. P, Jr., & Stenner, A. J. (2011). Integrating qualitative and quantitative research approaches via the phenomenological method. *International Journal of Multiple Research Approaches, 5*(1), 89–103.
Fisher, W. P, Jr., & Wilson, M. (2015). Building a productive trading zone in educational assessment research and practice. *Pensamiento Educativo: Revista de Investigacion Educacional Latinoamericana, 52*(2), 55–78.
Frodeman, R. (2014). Hermeneutics in the field: The philosophy of geology. In B. Babich & D. Ginev (Eds.), *The multidimensionality of hermeneutic phenomenology* (pp. 69–80). Heidelberg: Springer International Publishing.
Gadamer, H.G. (1970/2006) Language and understanding. *Theory, Culture & Society, 23*(1), 13–27.
Gadamer, H.-G. (1976). *Hegel's dialectic: Five hermeneutical studies* (P. C. Smith, Trans.). New Haven: Yale University Press.
Gadamer, H.-G. (1980). *Dialogue and dialectic: Eight hermeneutical studies on Plato* (P. C. Smith, Trans.). New Haven: Yale University Press.
Gadamer, H.-G. (1989). *Truth and method (J. Weinsheimer & D. G. Marshall, Trans.)* (Rev ed.). New York: Crossroad.
Gadamer, H.-G. (1991a). Gadamer on Gadamer. In H. J. Silverman (Ed.), *Continental Philosophy* (Vol. IV, pp. 13–19)., Gadamer and hermeneutics New York: Routledge.
Gadamer, H.-G. (1991b). *Plato's dialectical ethics: Phenomenological interpretations relating to the Philebus* (R. M. Wallace, Trans.). New Haven, Connecticut: Yale University Press.
Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago, Illinois: University of Chicago Press. ISBN 9780226279176.
Gasché, R. (2014). "A certain walk to follow:" Derrida and the question of method. Epoché. *A Journal for the History of Philosophy, 18*(2), 525–550.
Golinski, J. (2012). Is it time to forget science? Reflections on singular science and its history. *Osiris, 27*(1), 19–36.
Gooday, G. (1997). Instrumentation and interpretation: Managing and representing the working environments of Victorian experimental science. In B. Lightman (Ed.), *Victorian science in context* (pp. 409–437). Chicago: University of Chicago Press.
Hayek, F. A. (1948). *Individualism and economic order*. Chicago: University of Chicago Press.
Hayek, F. A. (1988). *The fatal conceit: The errors of socialism* (W. W. Bartley, III, Ed.) (Vol. I. The Collected Works of F. A. Hayek.) Chicago: University of Chicago Press.
Heelan, P. A. (1983). Natural science as a hermeneutic of instrumentation. *Philosophy of Science, 50*, 181–204.
Heelan, P. A. (1994). Galileo, Luther, and the hermeneutics of natural science. In T. J. Stapleton (Ed.), *The question of hermeneutics: Essays in honor of Joseph J. Kockelmans* (pp. 363–374). *Contributions to Phenomenology*, 17. Dordrecht, The Netherlands: Kluwer Academic Publishers.
Heelan, P. A. (2014). Consciousness, quantum physics, and hermeneutical phenomenology. In B. Babich & D. Ginev (Eds.), *The multidimensionality of hermeneutic phenomenology* (pp. 91–112). Heidelberg: Springer International Publishing.
Hegel, G. W. F. (1910/2003), *Phenomenology of mind*. Mineola, N.Y.: Dover Publications, Inc.
Heidegger, M. (1927/1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). New York: Harper & Row.
Heidegger, M. (1955/1991). *The principle of reason*, (R. Lilly, Trans.). Bloomington, Indiana: Indiana University Press.
Heidegger, M. (1975/1982). *The basic problems of phenomenology* (J. M. Edie, Ed.) (A. Hofstadter, Trans.). *Studies in Phenomenology and Existential Philosophy*. Bloomington, Indiana: Indiana University Press.

Heidegger, M. (1977). Modern science, metaphysics, and mathematics. In D. F. Krell (Ed.), *Basic writings* (pp. 243–282). New York: Harper & Row.

Heidtman, J., Wysienska, K., & Szmatka, J. (2000). Positivism and types of theories in sociology. *Sociological Focus, 33*(1), 1–26.

Husserl, E. (1911/1965). Philosophy as rigorous science (Q. Lauer, Trans.). In *Phenomenology and the crisis of philosophy* (pp. 69–147). New York: Harper & Row.

Husserl, E. (1913/1983). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy*. (F. Kersten, Trans.). Dordrecht: Kluwer.

Husserl, E. (1954/1970). *The crisis of European sciences and transcendental phenomenology: An introduction to phenomenological philosophy* (D. Carr, Trans.). Evanston, Illinois: Northwestern University Press.

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology.* The Indiana Series in the Philosophy of Technology). Bloomington, Indiana: Indiana University Press.

Kampen, J. K., & Tobi, H. (2011). Social scientific metrology as the mediator between sociology and socionomy In C. M. Baird (Ed.), *Social indicators: Statistics, trends and policy development* (pp. 1–26). New York, NY: Nova Science Publishers, Inc.

Kuhn, T. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.) *Criticism and the growth of knowledge* (pp. 91–195). Cambridge: Cambridge University Press.

Latour, B. (1986). Visualization and cognition: Thinking with eyes and hands. *Knowledge and Society: Studies in the Sociology of Culture Past and Present, 6*, 1–40.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. New York: Harvard University Press.

Latour, B. (1990). Postmodern? No, simply amodern! Steps towards an anthropology of science. *Studies in History and Philosophy of Science Part A, 21*(1), 145–171.

Latour, B. (1993). *We have never been modern. Translated by Catherine Porter*. Cambridge, Massachusetts: Harvard University Press.

Latour, B. (2005). *Reassembling the social: An introduction to Actor-Network-Theory*. Oxford, England: Oxford University Press.

Latour, B. (2013). *An inquiry into modes of existence* (C. Porter, Trans.). Cambridge, Massachusetts: Harvard University Press.

Martin, J., & Sugarman, J. (2001). Interpreting human kinds: Beginnings of a hermeneutic psychology. *Theory & Psychology, 11*(2), 193–207.

Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79*, 311–320.

McEvoy, J. G. (2007). Modernism, postmodernism and the historiography of science. *Historical Studies in the Physical and Biological Sciences, 37*(2), 383–408.

Michell, J. (2003). The quantitative imperative: Positivism, naïve realism and the place of qualitative methods in psychology. *Theory & Psychology, 13*(1), 5–31.

Moran, D. (2000). *Introduction to phenomenology*. London: Routledge.

Regan, P. (2012). Hans-Georg Gadamer's philosophical hermeneutics: Concepts of reading, understanding and interpretation. *Meta Research in Hermeneutics, Phenomenology and Practical Philosophy, 4*(2), 286–303.

Ricoeur, P. (1981). *Hermeneutics and the human sciences: Essays on language, action and interpretation* (J. B. Thompson, Ed. & Trans). Cambridge, England: Cambridge University Press.

Schieman, G. (2014). One cognitive style among others: Towards a phenomenology of the lifeworld and of other experiences. In B. Babich & D. Ginev (Eds.), *The multidimensionality of hermeneutic phenomenology* (pp. 310–348). Heidelberg: Springer International Publishing.

Shapere, D. (1984). *Reason and the search for knowledge: Investigations in the philosophy of science*. Dordrecht: D. Reidel Pub. Co.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations', and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science, 19*(3), 387–420.

Taylor, J. R., Cooren, F., Giroux, N., & Robichaud, D. (1996). The communicational basis of organization: Between the conversation and the text. *Communication Theory, 6*(1), 1–39.

Zammito, J. H. (2004). *A nice derangement of epistemes: Post-positivism in the study of science from Quine to Latour*. Chicago: The University of Chicago Press.

Measurement as a Medium for Communication and Social Action II: The Promise and Power of Being Amodern

William P. Fisher Jr. and Robert F. Cavanagh

A Challenge for the Social Sciences

As was elaborated in Part I, understanding the implications of an unmodern (Dewey 2012) or amodern (Latour 1999) philosophy for measurement, communication, and coordination begins from the way the use and development of language is not controlled or determined via a process of deliberate design. Hayek's (1948, p. 54) sense of this "central question of all social sciences" points toward the fundamentally symbolic nature of social reality, in Ricoeur's terms (1981, p. 219), and the active role of things themselves in conceptual formations not determined by any one thinker or actor (Gadamer 1989, p. 463). The technical media of language and communication, from phonemes to books to digital fonts to wifi to voltmeters, enable fragments of knowledge possessed by different people to spontaneously coordinate and self-organize in ways that may appear to have been centrally directed. But no individuals, organizations or governments have the depth or breadth of reach necessary for controlling the emergence or production of even phenomena as relatively simple as new slang terms, much less for the technical standards embodied in complex electronics or machines.

What are the conditions in which the spontaneous actions of individuals bring about distributions of resources that look as though they were made according to a single plan? Hayek's perspective on this problem of the social mind foregrounds a process of spontaneous coordinations (Birner and van Zijp 1994). How do those coordinations come about? We have a clue in Ricoeur's (1981, p. 193) observation that the reader's world horizon and opportunities for new directions and choices are

W.P. Fisher Jr. (✉)
University of California, Berkeley, CA, USA
e-mail: wfisher@berkeley.edu

R.F. Cavanagh
Curtin University, Bentley, WA, Australia
e-mail: R.Cavanagh@curtin.edu.au

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_11

expanded by new modes of being or forms of life received from a text. We all experience this appropriation of meaning and the projection of new world horizons from texts thousands of times a day. Reading a clock might tell me that I need to leave now to be at a meeting on time. Reading *Huckleberry Finn* might make aware of more opportunities for different moral choices in my life. Opening a window in response to a request for some fresh air might make work in a stuffy room more pleasant.

But past these mundane examples, we can see the general power of standards, for example, in the way that calibrating all clocks to the same measure of time simplifies the printing of train and plane schedules and the organization of meetings for people not in continuous contact. The efficiency of these coordinations is made possible by shared linguistic and technical standards embedded in thousands of word/concept/thing and instrument/theory/data assemblages. When we consider the combined effects of various standardized communications media, from pronunciation to grammar to vocabulary to symbols to technical specifications, we start to see how individuals' existential horizons come to be fused in a shared social world. In this regard, recent reconsiderations of operationalist and pragmatist philosophical perspectives in the context of psychological measurement offer productive directions for further inquiry (Maul et al. 2016).

Our immediate problem of interest is one of understanding how to coordinate local behaviors and decisions over a variety of different kinds of decisions across wider swaths of society in an uncoerced way that respects individual rights and liberties. Elaborating on the positive consequences of spontaneously self-organized effects, Hayek (1945, p. 88) quotes Whitehead's (1911, p. 61) observation that "Civilization advances by extending the number of important operations which we can perform without thinking about them." Indeed, everyday tools like telephones, computers, and automobiles are so complex that individual engineering experts do not have the range of knowledge needed to master all of the component parts in a single device. Most people have little more than the most elementary grasp of how their homes, furniture, clothing, or food are produced, and have even less of an inkling when it comes to their medications or their electronic communications, computing, and entertainment systems. Rudimentary levels of technical understanding do not, however, prevent the widespread use and enjoyment of an incredible range of appliances, tools, concepts and grammatical constructions, none of which could be created and deployed by any individual.

This phenomenon extends into science, where theoretical, experimental, and instrumental communities provide each other with ideas, results, and tools in uneven and disjointed fashion (Galison 1997, 1999). Instead of a positivist prioritization of observation, or the anti-positivist prioritization of theory, as the unifying focus of science, Galison offers a post-positivist intercalation of observation, theory, and instruments in an open-ended model that allows each area partial autonomy in effecting local coordinations of belief and action. The disunity of science in practice requires translations of ideas across these communities, none of which possesses all of the knowledge needed to produce useful results. The coordinations and

translations essential to alliances across areas of expertise counter-intuitively result in a stronger science than would be otherwise possible.

To take a simple example, Whitehead (1911, p. 62) reflects on the importance of a technology so seemingly unremarkable as Arabic numerals. The advantages of this symbol system relative to the Roman system are striking. When the Arabic numerals were complemented later by the introduction of symbols for arithmetic operations, the equals sign, and algebraic variables, a new kind of efficiency in the representation of relationships was secured. As Whitehead notes,

> Symbolization of the associative law (x+y=y+x), for instance, simplifies the representation of the idea that 'If a second number be added to any given number the result is the same as if the first given number had been added to the second number'. This example shows that, by the aid of symbolism, we can make transitions in reasoning almost mechanically by the eye, which otherwise would call into play the higher faculties of the brain. (p. 61)

This simple example of symbolic representations employed by billions of people daily illustrates the extent of everyone's dependency on networks of social conventions. No single individuals ever possess or create for themselves all, or even very much, of the information, knowledge, understanding, methods, or technology needed to achieve significant success in any area of life. Everyone is born and enculturated into a preexisting social world that provides a wealth of linguistic and technological tools essential to achievements of any kind. If every new person had to invent their own concepts, alphabets, phonemes, and grammar, and had to train others in their use, communication alone would be so time consuming that little else could be accomplished. Taking advantage of prior generations' products, from roads to institutions to familial norms, is our common human inheritance (Hyde 2010). Creative advancement takes place in dialogue with tradition, and often in opposition to it, but never in absence from it.

Mach (1919, p. 481) grasped that "Language, the instrument of this [scientific] communication, is itself an economical contrivance." That is, the economy of thought accomplished in symbolization is analogous to the economy of labor accomplished via machinery (Banks 2004). This economy is what Gadamer (1989, pp. 105, 428–430) refers to in terms of the way language prethinks the world for us, and so playfully lifts the burden of initiation by facilitating community. Hayek constructively identifies the spontaneous coordinations enacting the economy of thought as the central problem of all social sciences, and approvingly cites (Hayek 1948, p. 88) Whitehead's (1911) sense of advancing civilization by increasing the number of operations that can be performed without thinking about them. Hayek does not, however, follow through the implications of the insight to a full articulation of the role of technology (including texts, alphabets, and phonemes, in addition to tools and readable instruments) in an amodern account. Amodern perspectives on the coordination of word/concept/thing and instrument/theory/data assemblages across the various domains of experience have, however, become of intense interest in the philosophy, history, and social studies of the natural sciences, as much as in the social sciences and psychology (Ackermann 1985; Bud and Cozzens 1992; Dear 2012; Golinski 2012; Hutchins 1995, 2010, 2012, 2014;

Ihde 1991, 2012; Latour 1987, 2005, 2010, 2013). Indeed, despite having raised the question central to formulating an alternative amodern focus on the propagation of inscriptions across media, Hayek instead locates knowledge in the mind and ignores the common embodiment of knowledge in technology across fields.

Tarde (1902, 1903), in contrast, developed an entire economic psychology from the spontaneous coordinations of transactions at the individual level. Without espousing a superficial imitation of the natural sciences, Tarde held that success in quantifying individual variation in terms of relative importance, and not just in terms of what can easily be counted (Barry and Thrift 2007, p. 516; Tarde 1903, p. 107), would be an intellectual achievement of a new order surpassing the accomplishments of the natural sciences (Latour 2010). Ironically, quite in tune with Tarde's emphasis on the laws of social imitation, Hayek (1988, p. 21) held that "mind is not a guide but a product of cultural evolution, and is based more on imitation than on insight or reason."

But the central problem in tracing individuals' imitative patterns, for Tarde, is the lack of available instrumentation of this kind feeding the economic discipline and spreading out from the metrological chains (Latour 2010, p. 149; Latour and Lépinay 2010, p. 77). Where Hayek does not follow through on his insights and intuitions concerning the fundamental importance of technical embodiments of knowledge shared via social affiliations, Tarde, Latour, and the field of science and technology studies set the stage for the articulation of a full-fledged amodern paradigm of research and practice. Significant potential resides here for deepening and expanding the insights of Kahneman (2003) and others (Taleb 2012) concerning bounded rationality, the need to temper the overconfidence attending fast, intuitive, and emotional everyday reasoning, and the ways in which individuals' and groups' partial knowledge combine in a kind of stochastic resonance (Fisher and Wilson 2015, pp. 69–71). Extension of this paradigm into psychology and the social sciences is made possible by probabilistic measurement models and methods capable of contextualizing individual variation via metrological traceability.

Rasch models, especially in the construct mapping context (Stone et al. 1999; Wilson 2005, 2013) provide a means of achieving this kind of methodological authenticity because of the way they balance consideration of each facet of the theory/data/instrument assemblage (Fisher and Stenner 2011). As such, methods based in these models posit ideal uniformities, check observations against expectations, and focus attention on individual expressions of the learning progression, developmental sequence, or theoretical construct measured. The vast majority of Rasch model applications, however, remain caught up in data analysis, largely ignoring theory development and instrument calibration standards. Current practice explores the wide varieties of ways things come into presence, never arriving at or even pointing toward the possibility of systematic representations framing and exhibiting individual variations in a shared frame of reference. A new social conception of measurement theory and practice focuses on sorting out where and when it is possible for variations in the concrete ways individuals live out general forms of life to be meaningfully and usefully represented in common languages. A new metrology of psychometric, sociometric, and econometric unit standards could be

created by equating all instruments measuring the same thing, developing predictive theories capable of explaining causal relations (Stenner et al. 2013), and setting up traceability systems (Fisher 2009a, b; Fisher and Wilson 2015). Far from projecting unrealistic or inhumanly oppressive scientific ideals, this proposal seeks only to extend everyday model-based reasoning processes (Nersessian 1996, 2002, 2006, 2008, 2012, 2015) in new directions. We will now take up consideration of some practical consequences of this extension.

The Press of Externalities

Latour (1991, pp. 9–10) dismisses two facile clichés of hermeneutics often used to justify an ontological divide (as in Bryman 2007; Martin and Sugarman 2001) between the social and natural sciences. The first is that social scientists talk to people who talk back, while natural scientists talk to nature, which does not talk back. The second cliché is about the difference between the science of text interpretation and the inductive and deductive modes of reasoning applied to objects "out there" in the world. In dismissing the first cliché, Latour (1991) observes that natural scientists are as social as social scientists are in their interactions with each other, their publishers, their administrators, their funders, and their publics. As to the second, Latour (1991) says,

> walk to a laboratory, open a journal, talk with scientists, look at them: they are surrounded by hundreds of textual documents of various origins, of traces of different instruments, of faint parchments from decaying fossils, of subtle clues from more or less reliable polls; they are assembling them, reshuffling them, discounting some, stressing others…. Look at the exegetic work necessary to associate in a fine web iridium levels at the Cretaceous boundary, the dinosaurs' demise, the probability of meteorite impact, and nuclear winter…. No historian could be more astute in digging out an indefinite number of subtle clues and traces from archives than these natural scientists, who are doing just that, but in another literature.
>
> No, this exegetic work on faint and disparate traces is fully comparable to that of the scholars who establish the text of Plutarch out of twenty irreconcilable lessons, or those who reconstruct the daily occupations of the inhabitants of the caves of Lascaux or the split-second decay of the particles at CERN. Hermeneutics is not a characteristic of the social sciences; it is the property of all exegetic work; and, as far as texts are concerned, each department of any campus is made of exegetes who differ only in the source of their texts, not in the hermeneutic skill they deploy. All sciences are the offspring of biblical exegesis. The Book of Nature is the second tome of the Bible; this is what scientists since Galileo have echoed.

Hermeneutics, then, by Latour's own account, does much more than merely reinforce an artificial divide between the natural and social sciences. But Latour (1991, p. 16) rejects the notion of the hermeneutic as divisive without realizing the potential its universality brings to bear on matters of concern [adopting Latour's (2004) phrase here] in psychology and the social sciences. Latour's contributions to science and technology studies are significant (Mialet 2012), but this latter point, that "all sciences are the offspring of biblical exegesis," has, of course, been

explicitly developed in a large body of work not cited by Latour that extends at least from Heidegger's (1927, 1977a, b) studies of time, being, and technology, to Heelan's (1972, 1983, 1994, 1998) and Ihde's (1979, 1983, 1990, 1991, 1998; Ihde and Selinger 2003) explicitly hermeneutic and phenomenological concerns with instruments and sociotechnical contexts. Some of Heelan's and Ihde's work in this area predates Latour's earliest publications. As could be expected from the grasp of the four characteristics of phenomenology (Fisher and Cavanagh 2016) evident in their works, their recent efforts (Heelan 2003, 2013; Ihde 1991, 2009, 2012) expand on themes shared with Latour, overlapping to some degree with, but also differing from, the work of Latour's colleague, Harman (2002, 2005a, b, 2009), on Heidegger.

It is true that what Latour (1995) calls the propagation of inscriptions across media ought indeed be as much a focus of interest in the social sciences as it is the natural. But apart from Heelan's (1993, 1995) reflections on the philosophical implications of quantum mechanics for the social sciences, none of these writers attempts to bring the consequences of an amodern, post-positivist perspective to bear on measurement in psychology and the social sciences. Conversely, the recognition of language as medium in Gadamer's and Ricoeur's hermeneutics is left undeveloped by them as to the implications for measuring instruments as texts written and read in common symbol systems. They fail to follow through from (a) their recognitions of the universality of the fact that all objectification and explanatory processes take place within a sphere of standardized signs to (b) the networks of distributed technologies and associated imitative behaviors that extend and embody those signs' representations.

And so it happens that questions concerning potentials for framing and revealing individual differences via common languages, shared standards, and more efficient information markets in psychology and the social science remain largely unasked and unexplored. Most work addressing the role of standards in reducing information transaction costs (Barzel 1982; Swann 2005; Weitzel 2004) simply ignores the potentials improved psychological and social measurement offer the economics of human, social and natural capital (Fisher 2010). In one particularly pointed example, for instance, Latour and Callon (2011) explicitly address how the natural sciences are able to coordinate the capital formatting operations that make markets efficient, while assuming no parallel or analogous kinds of coordinations will ever be possible in the social sciences. Instead, the formatting operations of the social sciences are conceived as capable only of incessantly producing market externalities. Most telling is that the reasons for this are not understood as being located in the quality of the representations produced and the lack of metrological traceability for the putative quantities. No consideration at all is given to the possibility that human, social, and natural capital could ever surmount the "incommensurable difference that divides the internalities from the externalities" (Latour and Callon 2011, p. 20). One of several key aspects of the larger economic problem (Fisher 2011, 2012b) is, however, rightly identified by Latour and Callon as ownership: "the constant effort to internalize the externalities cannot be successful unless a previous sensible distribution of property rights has taken place" (p. 7). Those

rights, of course, depend on clear, simple, uniform, and universally accessible ways of knowing the amount, quality, and price of the capital stock in question.

Unfortunately, Latour and Callon do not address the idea that intangible assets could be viewed productively as resources helping economics achieve its "main goal—the creation of calculable and governable spaces through the production of internalities." Hayek's (1988) focus on the "fatal conceit" of socialism's Cartesian subject, and recent examinations of the roles of institutions in the creation of markets (Miller and O'Leary 2007), both come to bear here. Internalizing the externalities will require new ways of approaching the contrast between networks' spontaneously coordinated determinations of amount, quality and price, on the one hand, and centrally administered control over those determinations, on the other. Economic models internalizing human, social, and natural forms of capital have long since been proposed (Ekins 1992). The gift character of the objects of the social sciences—the spontaneous self-organization of their invariant constructs and the opportunities presented for transparent, additive, mobile, and divisible representations—remains completely invisible to anyone who is not explicitly looking for it. The invention of legal rights, metrological standards, traceability and quality assurance systems, and accounting rules supporting the ownership of intangible assets appear necessary to overcome the abuses and inequities occurring when economically internalized private property is restricted to manufactured capital and land. These rights, standards, systems, and rules would not expand the power of the state, but would on the contrary empower the emergence of new markets that reduce its ability to satisfy vested interests by means of economically inefficient actions.

The theory and practice of metrological traceability in psychological and social measurement have expanded beyond its origins in methods of instrument equating and item banking (Rasch 1960; Andrich 1988, 2004; Bond and Fox 2015; Engelhard 2012; Jolander 2008; Wilson 2005; Wright 1977, 1999). Explorations of possibilities for unit standards have demonstrated their viability in experimental, instrumental and theoretical terms (Fisher 1997, 1999a, b, 2000, 2005, 2009a, 2012a, b; Fisher et al. 1995). These possibilities have more recently given rise to psychometric–engineering partnerships in working out a common language (Mari and Wilson 2013; Pendrill 2014; Pendrill and Fisher 2013, 2015; Wilson et al. 2015). From here, future alliances and others already in progress will bring measurement stakeholder groups already using high-quality measurement into alignment with other groups key to practicing the "art of interessement" (Akrich et al. 2002). Significant numbers of psychometricians, engineers, educators, health care researchers, and quality improvement specialists, along with smaller numbers of business managers (Drehmer et al. 2000; Ewing et al. 2005; Salzberger 2009; Salzberger and Sinkovics 2006), and IT developers (Drehmer and Deklava 2001; Torres Irribarra et al. 2015), are aware of the advantages of attending to measurement details, but their network of alliances has so far been insufficient to shifting the paradigm. That may change, and critical mass might be reached, as other groups invested in high quality measurement, such as the Sustainability Accounting Standards Board (Gleeson-White 2015), forensic metrologists

(Vosk 2010, 2013; Vosk and Emery 2015), legal metrologists and others, recognize the viability of new obligatory passage points shared with other select allies.

Further, another idea lost on Latour and Callon (2011) is that the critique of capitalism is itself capitalist, just as the critique of science must itself be scientific. Socialism failed because it could not organize capital more efficiently than capitalism, but huge potential for improving not just the efficiency of capital markets but their moral bearing in the world nonetheless remains. This is especially so for human, social, and natural capital markets. Vast improvements are possible, and contrary to Latour and Callon's view, the unification of capitalism does not necessarily entail dehumanizing reductions to quantities useful only to centralized manipulations and control of human behavior and associations. They are quite right to point at the need to research the history of science, technology, and metrology, but they do not see the opportunities we have for using the results of that study to stop the externalizations and start internalizing the objects of the social sciences in ways that could productively reduce and reverse human suffering, social discontent, and environmental degradation (Fisher 2009a, b, 2011, 2012a, b).

That is, given the four characteristics of phenomenology in Paper One, can we formulate a provisional answer to Latour and Callon's (2011, p. 20) question, "how is one to take charge of the externalities incessantly produced by the formatting machines of the social sciences?" Contrary to what Latour and Callon (p. 22) expect, might there be a way in which the social sciences can take charge of the externalities and participate in their inscription and internalization in a way that is not inhumanly reductionistic, and socially and environmentally destructive? Might there be an opportunity for making profit contingent on growth in well managed and measured human, social, and natural capital? If each species (form of life, mode of being) of actant (literacy, health, functionality, etc.) was situated in its special ecological niche of relationships (mechanically put: linked in metrological chains), the assumption that profits are only ever enhanced at the expense of human, social and natural capital might well be proven mistaken. Positively restating that we can try to imagine how investments in human, social, and natural capital might be translated into scientifically, legally, and financially accountable returns, instead of simply assuming such accounting could never be possible.

Might a theory and practice of genuine wealth (Ekins 1992, 1999; Ekins et al. 2008) replace the dismal science of economics focused only on the measurement and management of manufactured capital and property? Must the legal, accounting, psychometric, classroom, IT, clinical, regulatory, research, etc., implementations of statistical methods be forever doomed to manipulations of incommensurable, locally dependent numbers assumed to stand for constructs, but never demonstrated empirically or theoretically as doing so? Longstanding methods of construct mapping, instrument equating, item banking, and theory development have already been used to create psychological and social metrological chains enduring for decades. Might these finally be systematically and creatively applied to gather all stakeholders as allies around the same boundary object relative to agreed-upon

obligatory passage points and standards? Should not we extend to human, social, and natural capital the institutional rules, roles and responsibilities already enabling the alignment and coordination of investments in manufactured capital?

Amodern Measurement Theory

The four characteristics of phenomenology in Paper One foreground the requirements of amodern measurement theory. This paper transitions from that foregrounding to an amodern account of how the world of the text opens onto a new answer to Hayek's question concerning the spontaneous coordination of decisions based on partial knowledge.

The unity of subject and object embodied together in a shared technological medium is of fundamental consequence to amodern measurement theory and practice. The universal importance of this amodern perspective is evident in that, "In quantum mechanics as in psychiatry, in ecology as much as in anthropology, the scientific observer is now—willy-nilly—also a *participant*" (Toulmin 1982, p. 97). Participant observation entails an ecology of mind (Bateson 1972) embodied in linguistic media (Hutchins 2012) providing each major species of actant a way to assert its mode of being as a form of life in the network of associations.

Practical applications of this amodern sense of measurement require the integration of individual-level agent-based models of evolving ecological relationships (Epstein 1999; Grimm and Railsback 2013; Jolly and Wakeland 2011; Railsback 2001), hierarchical linear models (Bryk and Raudenbush 1992; Kamata 2001), and Rasch's models for measurement (Rasch 1960). Each class of models brings specific features to bear that are necessary to an amodern measurement paradigm.

Agent-based and neural net models in ecological and economic research provide a means for simulating, planning, and innovating across a wide range of actant interrelations, from the natural to the social and all combinations of them. Applications of these models remain oriented largely in a modern perspective toward data analysis and centralized control, and not toward an amodern approach to designing and deploying ecosystems of individual actors equipped with the tools they need to survive and thrive in their local niches. Rasch models facilitate the latter, as well as other features lacking in agent-based models, such as minimally sufficient statistics, and calibrated representations of relationships expressed in common languages enabling coordinated practical applications for end users.

Combining agent-based and hierarchical linear models will enable close study of multilevel ecologies of relationships within, between, and among individuals of any given species of actants. Integrating these with Rasch models in practical applications opens the door to a new paradigm of evolving adaptive ecological complexity in psychology and the social sciences.

Conclusion

Amodern measurement in the social sciences extends beyond quantifying specific qualities of an individual to the development and deployment of technical communications systems representing interactions between individuals and between individuals and their worlds. Amodern measurement coordinates the collecting and dissemination of information through metrological networks and processes. Early stage research tests for correspondences between theory, data, and instruments that follow from the way independent objects of investigation assert themselves as reproducible and real. Given consistent results, the agent compelling agreement among observers as to its separate existence as a thing in the world is then transformed into a product of agreement conceptually situated in an ecology of relationships and marked by a standardized nomenclature (Fisher 2000; Ihde 1991; Latour 1987, 2005; Wise 1995). Amodern measurement is therefore concerned with structuring collaboration and communication between many measurers and developing common understandings of what is being measured in distributed networks. This is a move away from the construction and application of instruments in isolation toward a convergent sharing of theoretical knowledge and data from many instruments. The vehicle for communication is the semiotic combination of the technological medium (a word, a measuring instrument, etc.), the thing measured (time, mass, reading ability, health, etc.), and the conceptual, theoretical meaning of the ideas involved.

These two papers were structured around two related themes that coalesced to describe a new theory of measurement. The philosophical foundation of the theory is phenomenology, in particular four characteristics of phenomenology in conjunction with a measured dismissal of traditional conceptions of science including modernism. Having never actually been modern, insofar as science has not ever been able to put the positivist assumptions of modern Cartesianism into actual practice, it makes no sense to try to be postmodern (Latour 1991, 1993). "Postmodernism is a disappointed form of modernism" (Latour 1991, p. 17). A far better strategy, taken up in different ways by Glazebrook (2000), Harman (2002, 2005b), Heelan (1994), Ihde (1997) and others, is to follow Heidegger's (1959) efforts in trying to revive the concepts of ancient Greek pre-modern thinking as a complete alternative to modernism, one not defined in relation to it. Amodern measurement theory's incorporation of unified subject-objects conceived as ecologically interrelated and evolving forms of life offers a way forward that avoids the limitations inherent in notions of positivism and anti-positivism. Significantly, key constructs in the theory have application in evaluating the practice of measurement in contemporary human science research and in the signaling of future directions, particularly as concerns the integration of agent-based ecological models, Rasch models, and hierarchical linear models.

References

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science*. Princeton, New Jersey: Princeton University Press.

Akrich, M., Callon, M., & Latour, B. (2002). The key to success in innovation. Part I: The art of interessement. *International Journal of Innovation Management, 6*(2), 187–206.

Andrich, D. (1988). *Rasch models for measurement.* (Vols. series no. 07-068). (Sage University Paper Series on Quantitative Applications in the Social Sciences). Beverly Hills, California: Sage Publications.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), I-7–I-16.

Banks, E. (2004). The philosophical roots of Ernst Mach's economy of thought. *Synthese, 139*(1), 23–53.

Barry, A., & Thrift, N. (2007). Gabriel tarde: Imitation, invention and economy [introduction to a special issue on G. Tarde]. *Economy and Society, 36*(4), 509–525.

Barzel, Y. (1982). Measurement costs and the organization of markets. *Journal of Law and Economics, 25*, 27–48.

Bateson, G. (1972). *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*. University of Chicago Press.

Birner, J., & van Zijp, R. (1994). *Hayek, coordination and evolution*. New York: Routledge.

Bond, T., & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, California: Sage Publications.

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research, 1*(1), 8–22.

Bud, R., & Cozzens, S. E. (Eds). (1992). *SPIE Institutes. Vol. 9: Invisible connections: Instruments, institutions, and science* (R. F. Potter, Ed.). Bellingham, WA: SPIE Optical Engineering Press.

Dear, P. (2012). Science is dead; long live science. *Osiris, 27*(1), 37–55.

Dewey, J. (2012). *Unmodern philosophy and modern philosophy* (P. Deen, Ed.). Carbondale, Illinois: Southern Illinois University Press.

Drehmer, D. E., & Deklava, S. M. (2001). A note on the evolution of software engineering practices. *Journal of Systems and Software, 57*(1), 1–7.

Drehmer, D. E., Belohlav, J. A., & Coye, R. W. (2000). A exploration of employee participation using a scaling approach. *Group and Organization Management, 25*(4), 397–418.

Ekins, P. (1992). A four-capital model of wealth creation. In P. Ekins & M. Max-Neef (Eds.), *Real-life economics: Understanding wealth creation* (pp. 147–155). London: Routledge.

Ekins, P. (1999). *Economic growth and environmental sustainability: The prospects for green growth*. New York: Routledge.

Ekins, P., Dresner, S., & Dahlstrom, K. (2008). The four-capital method of sustainable development evaluation. *European Environment, 18*(2), 63–80.

Engelhard, G, Jr. (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge Academic.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Generative Social Science: Studies in Agent-Based Computational Modeling, 4*(5), 4–46.

Ewing, M. T., Salzberger, T., & Sinkovics, R. R. (2005). An alternative approach to assessing cross-cultural measurement equivalence in advertising research. *Journal of Advertising, 34*(1), 17–36.

Fisher, W. P, Jr. (1997). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement, 1*(2), 87–113.

Fisher, W. P, Jr. (1999a). Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. *Journal of the Louisiana State Medical Society, 151*(11), 566–578.

Fisher, W. P, Jr. (1999b). Metrology note. *Rasch Measurement Transactions, 13*(3), 774.

Fisher, W. P, Jr. (2000). Objectivity in psychosocial measurement: What, why, how. *Journal of Outcome Measurement, 4*(2), 527–563.

Fisher, W. P, Jr. (2005). Daredevil barnstorming to the tipping point: New aspirations for the human sciences. *Journal of Applied Measurement, 6*(3), 173–179.

Fisher, W. P, Jr. (2009a). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement, 42*(9), 1278–1287.

Fisher, W. P., Jr. (2009b). *NIST Critical national need idea White Paper: metrological infrastructure for human, social, and natural capital* (Tech. Rep., http://www.nist.gov/tip/wp/pswp/upload/202_metrological_infrastructure_for_human_social_natural.pdf). Washington, DC: National Institute for Standards and Technology.

Fisher, W. P., Jr. (2010). *Measurement, reduced transaction costs, and the ethics of efficient markets for human, social, and natural capital.* Bridge to Business Postdoctoral Certification, Freeman School of Business, Tulane University. http://ssrn.com/abstract=2340674.

Fisher, W. P, Jr. (2011). Bringing human, social, and natural capital to life: Practical consequences and opportunities. In N. Brown, B. Duckor, K. Draney, & M. Wilson (Eds.), *Advances in Rasch Measurement* (Vol. 2, pp. 1–27). Maple Grove, MN: JAM Press.

Fisher, W. P, Jr. (2012a). Measure and manage: Intangible assets metric standards for sustainability. In J. Marques, S. Dhiman, & S. Holt (Eds.), *Business administration education: Changes in management and leadership strategies* (pp. 43–63). New York: Palgrave Macmillan.

Fisher, W. P., Jr. (2012b). What the world needs now: A bold plan for new standards [Third place, 2011 NIST/SES World Standards Day paper competition]. *Standards Engineering, 64*(3), 1 & 3–5 [http://ssrn.com/abstract=2083975].

Fisher, W. P., Jr., & Cavanagh, R. F. (2016). *Measurement: A medium for communication and social action:* I. *A phenomenological view of science and society*. Paper presented at the 2015 Annual Symposium of the Pacific Rim Objective Measurement Symposium, August 4–6, Fukuoka. Heidelberg: Springer.

Fisher, W. P, Jr., & Stenner, A. J. (2011). Integrating qualitative and quantitative research approaches via the phenomenological method. *International Journal of Multiple Research Approaches, 5*(1), 89–103.

Fisher, W. P, Jr., & Wilson, M. (2015). Building a productive trading zone in educational assessment research and practice. *Pensamiento Educativo: Revista de Investigacion Educacional Latinoamericana, 52*(2), 55–78.

Fisher, W. P, Jr., Harvey, R. F., Taylor, P., Kilgore, K. M., & Kelly, C. K. (1995). Rehabits: A common language of functional assessment. *Archives of Physical Medicine and Rehabilitation, 76*(2), 113–122.

Gadamer, H.-G. (1989). *Truth and method (J. Weinsheimer & D. G. Marshall, Trans.)* (Rev ed.). New York: Crossroad.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago, Illinois: University of Chicago Press.

Galison, P. (1999). Trading zone: Coordinating action and belief. In M. Biagioli (Ed.), *The science studies reader* (pp. 137–160). New York: Routledge.

Glazebrook, T. (2000). *Heidegger's philosophy of science*. New York: Fordham University Press.

Gleeson-White, J. (2015). *Six capitals, or can accountants save the planet? Rethinking capitalism for the 21st century*. New York: Norton.

Golinski, J. (2012). Is it time to forget science? Reflections on singular science and its history. *Osiris, 27*(1), 19–36.

Grimm, V., & Railsback, S. F. (2013). *Individual-based modeling and ecology*. Princeton, NJ: Princeton University Press.

Harman, G. (2002). *Tool-being: Heidegger and the metaphysics of objects*. Chicago: Open Court.
Harman, G. (2005a). *Guerrilla metaphysics: Phenomenology and the carpentry of things*. Chicago: Open Court.
Harman, G. (2005b). Heidegger on objects and things. In B. Latour & P. Weibel (Eds.), *Making things public: Atmospheres of democracy* (pp. 268–271). Cambridge, MA: MIT Press.
Harman, G. (2009). *Prince of networks: Bruno Latour and metaphysics*. Melbourne, Australia: Re.press.
Hayek, F. A. (1948). *Individualism and economic order*. Chicago: University of Chicago Press.
Hayek, F. A. (1988). *The fatal conceit: The errors of socialism* (W. W. Bartley, III, Ed.) (Vol. I). The Collected Works of F. A. Hayek. Chicago: University of Chicago Press.
Heelan, P. A. (1972). Hermeneutics of experimental science in the context of the lifeworld. *Philosophia Mathematica, s1–9,* 101–144.
Heelan, P. A. (1983). Natural science as a hermeneutic of instrumentation. *Philosophy of Science, 50*, 181–204.
Heelan, P. A. (1993). *Theory of social-historical phenomena: Quantum mechanics and the social sciences.* In *Continental Philosophy and Social Science.* Society for the Philosophy of the Human Sciences, with the Society for Phenomenology and Existential Philosophy, New Orleans, LA.
Heelan, P. A. (1994). Galileo, Luther, and the hermeneutics of natural science. In T. J. Stapleton (Ed.), *The question of hermeneutics: Essays in honor of Joseph J. Kockelmans* (pp. 363–374). *Contributions to Phenomenology*, 17. Dordrecht, The Netherlands: Kluwer Academic Publishers.
Heelan, P. A. (1995). Quantum mechanics and the social sciences: After hermeneutics. *Science & Education, 4*(2), 127–136.
Heelan, P. A. (1998). The scope of hermeneutics in natural science. *Studies in History and Philosophy of Science Part A, 29*(2), 273–298.
Heelan, P. A. (2003). Husserl, Lonergan, and paradoxes of measurement. *Journal of Macrodynamic Analysis, 3*, 76–96.
Heelan, P. A. (2013). Phenomenology, ontology, and quantum physics. *Foundations of Science, 18*(2), 379–385. doi:10.1007/s10699-011-9247-6.
Heidegger, M. (1927/1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). New York: Harper & Row.
Heidegger, M. (1953/1959). *An introduction to metaphysics* (R. Manheim, Trans.). New Haven, Connecticut: Yale University Press.
Heidegger, M. (1977a). Modern science, metaphysics, and mathematics. In D. F. Krell (Ed.), *Basic writings* (pp. 243–282). New York: Harper & Row.
Heidegger, M. (1977b). The question concerning technology. In D. F. Krell (Ed.), *Basic writings* (pp. 283–317). New York: Harper & Row.
Hutchins, E. (1995). *Cognition in the wild*. Cambridge, Massachusetts: MIT Press.
Hutchins, E. (2010). Cognitive ecology. *Topics in Cognitive Science, 2*, 705–715.
Hutchins, E. (2012). Concepts in practice as sources of order. *Mind, Culture, and Activity, 19*, 314–323.
Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philosophical Psychology, 27*(1), 34–49.
Hyde, L. (2010). *As common as air: Revolution, art, and ownership*. New York: Farrar, Straus and Giroux.
Ihde, D. (1979). *Technics and praxis: A philosophy of technology*. Dordrecht, Holland: D. Reidel.
Ihde, D. (1983). The historical and ontological priority of technology over science. *Existential technics* (pp. 25–46). Albany, New York: State University of New York Press.
Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Bloomington, Indiana: Indiana University Press.
Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology. The Indiana Series in the Philosophy of Technology*. Bloomington, Indiana: Indiana University Press.

Ihde, D. (1997). Thingly hermeneutics/technoconstructions. *Man and World, 30*(3), 369–381.
Ihde, D. (1998). *Expanding hermeneutics: Visualism in science*. Evanston, Illinois: Northwestern University Press.
Ihde, D. (2009). *Postphenomenology and technoscience: The Peking University lectures*. Stony Brook, New York: State University of New York Press.
Ihde, D. (2012). *Experimental phenomenology: Multistabilities* (2nd ed.). Albany, New York: SUNY Press.
Ihde, D., & Selinger, E. (Eds.). (2003). *Chasing technoscience: Matrix for materiality*. Bloomington, Indiana: Indiana University Press.
Jolander, F. (2008). Something about bridge-building [test-equating] techniques–a sensational new creation by Dr. Rasch (C. Kreiner, Trans.). *Rasch Measurement Transactions 21*(4): 1129–1130 [http://www.rasch.org/rmt/rmt214a.htm]. (Reprinted from *Folkeskolen [The Danish Elementary School Journal*], 23 May 1957).
Jolly, R., & Wakeland, W. (2011). Using agent based simulation and game theory analysis to study knowledge flow in organizations: The KMscape. In M. Jennex (Ed.), *Global aspects and cultural perspectives on knowledge management: Emerging dimensions* (pp. 19–29). Hershey, PA: Information Science Reference.
Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79–93.
Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*(5), 1449–1475.
Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. New York: Harvard University Press.
Latour, B. (1991). The impact of science studies on political philosophy. *Science, Technology and Human Values, 16*(1), 3–19.
Latour, B. (1993). *We have never been modern*. Cambridge, Massachusetts: Harvard University Press
Latour, B. (1995). Cogito ergo sumus! Or psychology swept inside out by the fresh air of the upper deck: Review of Hutchins' Cognition in the Wild, MIT Press, 1995. *Mind, Culture, and Activity: An International Journal, 3*(192), 54–63.
Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge, Massachusetts: Harvard University Press.
Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. *Critical Inquiry, 30*(2), 225–248.
Latour, B. (2005). *Reassembling the social: An introduction to Actor-Network-Theory.* (Clarendon Lectures in Management Studies). Oxford, England: Oxford University Press.
Latour, B. (2010). Tarde's idea of quantification. In M. Candea (Ed.), *The social after Gabriel tarde: Debates and assessments* (pp. 145–162). London: Routledge.
Latour, B. (2013). *An inquiry into modes of existence* (C. Porter, Trans.). Cambridge, Massachusetts: Harvard University Press.
Latour, B., & Callon, M. (2011). "Thou shall not calculate!" or how to symmetricalize gift and capital. *Revista De Pensamiento e Investifation Social, 11*(1), 171–192.
Latour, B., & Lépinay, V. A. (2010). *The science of passionate interests: An introduction to Gabriel Tarde's economic anthropology*. Prickly Paradigm Press.
Mach, E. (1919). *The science of mechanics: A critical and historical account of its development (T. J. McCormack, Trans.)* (4th ed.). Chicago: The Open Court Publishing Co. (Original work published 1883).
Mari, L., & Wilson, M. (2013). A gentle introduction to Rasch measurement models for metrologists. *Journal of Physics Conference Series, 459*(1). http://iopscience.iop.org/1742-6596/459/1/012002/pdf/1742-6596_459_1_012002.pdf.
Martin, J., & Sugarman, J. (2001). Interpreting human kinds: Beginnings of a hermeneutic psychology. *Theory & Psychology, 11*(2), 193–207.
Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79*, 311–320.

Mialet, H. (2012). Where would STS be without Latour? What would be missing? *Social Studies of Science, 42*(3), 456–461.
Miller, P., & O'Leary, T. (2007). Mediating instruments and making markets: Capital budgeting, science and the economy. *Accounting, Organizations and Society, 32*(7–8), 701–734.
Nersessian, N. J. (1996). Child's play. *Philosophy of Science, 63*, 542–546.
Nersessian, N. J. (2002). Maxwell and "the method of physical analogy": Model-based reasoning, generic abstraction, and conceptual change. In D. Malament (Ed.), *Reading natural philosophy: Essays in the history and philosophy of science and mathematics* (pp. 129–166). Lasalle, Illinois: Open Court.
Nersessian, N. J. (2006). Model-based reasoning in distributed cognitive systems. *Philosophy of Science, 73*, 699–709.
Nersessian, N. J. (2008). *Creating scientific concepts*. Cambridge, Massachusetts: MIT Press.
Nersessian, N. J. (2012). Engineering concepts: The interplay between concept formation and modeling practices in bioengineering sciences. *Mind, Culture, and Activity, 19*, 222–239.
Nersessian, N. J. (2015). Conceptual innovation on the frontiers of science. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 455–474). Cambridge, MA: MIT Press.
Pendrill, L. (2014). Man as a measurement instrument [Special Feature]. *NCSLi Measure: The Journal of Measurement Science, 9*(4), 22–33.
Pendrill, L., & Fisher, W. P., Jr. (2013). Quantifying human response: Linking metrological and psychometric characterisations of man as a measurement instrument. *Journal of Physics: Conference Series, 459,* http://iopscience.iop.org/1742-6596/459/1/012057.
Pendrill, L., & Fisher, W. P, Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement, 71*, 46–55.
Railsback, S. F. (2001). Concepts from complex adaptive systems as a framework for individual-based modelling. *Ecological Modelling, 139*(1), 47–62.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.
Ricoeur, P. (1981). *Hermeneutics and the human sciences: Essays on language, action and interpretation* (J. B. Thompson, Ed. & Trans). Cambridge, England: Cambridge University Press.
Salzberger, T. (2009). *Measurement in marketing research: An alternative framework*. Northampton, MA: Edward Elgar.
Salzberger, T., & Sinkovics, R. R. (2006). Reconsidering the problem of data equivalence in international marketing research: Contrasting approaches based on CFA and the Rasch model for measurement. *International Marketing Review, 23*(4), 390–417.
Stenner, A. J., Fisher, W. P, Jr., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology: Quantitative Psychology and Measurement, 4*(536), 1–14.
Stone, M. H., Wright, B., & Stenner, A. J. (1999). Mapping variables. *Journal of OutcomeMeasurement, 3*(4), 308–322.
Swann, G. M. P. (2005, 2 December). John Barber's pioneering work on the economics of measurement standards [Electronic version]. Retrieved http://www.cric.ac.uk/cric/events/jbarber/swann.pdf from Notes for Workshop in Honor of John Barber held at University of Manchester.
Taleb, N. N. (2012). *Antifragile: Things that gain from disorder*. New York: Random House.
Tarde, G. (1902). *Psychologie economique*. Paris: Ancienne Librairie Germer Baillière et Cie.
Tarde, G. (1903). Inter-psychology, the inter-play of human minds. *International Quarterly, 7*, 59–84.
Torres Irribarra, D., Freund, R., Fisher, W. P., Jr., & Wilson, M. (2015). Metrological traceability in education: A practical online system for measuring and managing middle school mathematics instruction. *Journal of Physics Conference Series, 588*, http://iopscience.iop.org/1742-6596/588/1/012042.

Toulmin, S. E. (1982). The construal of reality: Criticism in modern and postmodern science. *Critical Inquiry, 9*, 93–111.

Vosk, T. (2010). Trial by numbers: Uncertainty in the quest for truth and justice. *The Champion (National Association of Criminal Defense Lawyers), 56*, 48–56.

Vosk, T. (2013). Measurement uncertainty. In *Encyclopedia of Forensic Sciences, Second Edition* (Vol. 3, pp. 322–331). Waltham, MA: Academic Press.

Vosk, T., & Emery, A. F. (2015). *Forensic metrology: Scientific measurement and inference for lawyers, judges and criminalists* (M. Houk, Ed.). International Forensic Science and Investigation Series. Boca Raton, FL: CRC Group, Taylor & Francis.

Weitzel, T. (2004). *Economics of standards in information networks*. New York: Physica-Verlag.

Whitehead, A. N. (1911). *An introduction to mathematics*. New York: Henry Holt and Co.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wilson, M. R. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement, 46*, 3766–3774.

Wilson, M., Mari, L., Maul, A., & Torres Irribarra, D. (2015). A comparison of measurement concepts across physical science and social science domains: Instrument design, calibration, and measurement. *Journal of Physics: Conference Series, 588*(012034), http://iopscience.iop.org/1742-6596/588/1/012034.

Wise, M. N. (1995). Precision: Agent of unity and product of agreement. Part III–"Today Precision Must Be Commonplace." In M. N. Wise (Ed.), *The values of precision* (pp. 352–61). Princeton, New Jersey: Princeton University Press.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97–116.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 65–104). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

A Hyperbolic Cosine Unfolding Model for Evaluating Rater Accuracy in Writing Assessments

Jue Wang and George Engelhard Jr.

Murphy and Cleveland (1995) defined rating accuracy as one type of rating quality index, which has been used to evaluate rating quality, rater performance, and even training efficiency. Engelhard (2013) indicated that "accuracy can be defined as the comparison between unknown processes and a defined standard process in order to adjust the unknown process until it matches the standard" (p. 232). In other words, accuracy measures the distance between the operational ratings and the criterion ratings. There are many ways to obtain the criterion ratings. Engelhard (1996) used the criterion ratings defined by a panel of experts. Wolfe and McVay (2012) have defined the true ratings based on average ratings across the operational raters. Rater accuracy within the context of writing assessment can be defined as a latent trait indicating how accurate a rater is in scoring the essays (Engelhard 2013). Other ways to calculate accuracy ratings have been proposed (Cronbach 1955; Sulsky and Balzer 1988; Jones 2007).

Unfolding models were developed and widely used in the context of attitude measurement (Thurstone 1927, 1928; Thurstone and Chave 1929; Coombs 1964; Coombs and Avrunin 1977; Poole 2005; Davison 1977). Probabilistic unfolding models have been proposed by several researchers (Andrich 1988, 1995; Luo 1998, 2001; Roberts and Laughlin 1996; Roberts et al. 2002). Unfolding models can differentiate the directionality of the disagree category by unfolding the responses into two categories-Disagree Below and Disagree Above. Among the family of unfolding models (Luo and Andrich 2005), the Hyperbolic Cosine unfolding model (HCM) has good properties for examining rater accuracy because a unit parameter that can be estimated as a property of the data is independent of the scale. This unit parameter also has interesting substantive interpretation that can help understand rater accuracy as will be described below. Therefore, HCM is chosen for this study

J. Wang (✉) · G. Engelhard Jr.
The University of Georgia, 325W Aderhold Hall, 110 Carlton Street,
Athens, GA 30602, USA
e-mail: cherish@uga.edu

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_12

to evaluate rater accuracy within the context of writing assessment. The HCM is described in the following section.

During rater training and scoring, raters are usually asked to study the rating rubrics and benchmarks essays. These detailed rating rubrics indicate what an essay should look like in order to obtain a certain level of score either holistic or domain-based. For example, a rubric describing the domain of *Organization* for a rating of 3 and 4 could be labeled as *ideas are partial developed* and *generally well-developed ideas* for the rating rubrics. The judgmental processes of each rater are unique, so the cognitive processes of raters on each domain of different essays are difficult to control. Even though we would like all the raters to be fair and consistent, we also recognize that as independent-thinking human beings raters are different from each other, for example between more and less proficient raters. If a rater is biased and inaccurate, then the observed rating score may be higher or lower than the criterion ratings. Therefore, for a domain, such as *Organization* of essays, inaccurate responses of raters may be due to two totally opposite reasons. Using Rasch models based on cumulative response processes, we can only get part of the information that whether raters provide accurate or inaccurate ratings for *Organization*. Unfolding models can differentiate the directionality of inaccuracy, and they offer possibilities to evaluate essay characteristics and rater cognition by having additional information.

The purpose of this study is to evaluate rater accuracy on a statewide writing assessment using the HCM. A secondary purpose is to explore the usefulness of unfolding models by developing substantive interpretation related to rater judgments for parameters in the HCM.

Comparing Cumulative Process and Unfolding Process

Cumulative data follow a Guttman pattern, and it is well known that the Rasch model implies a probabilistic Guttman pattern in item response data. Table 1 shows seven stimuli (mathematics items) ordered from easy to hard where a person who answered a harder item correctly should also get the easier items correct. However, the unfolding data follow a parallelogram pattern. In Table 1, we have seven stimuli measuring the preference towards cats, and we order them from dislike to like, then a group of persons who responded *Yes* for Stimulus G is very unlikely to agree with the statements A and B. Similarly, a group of persons who chose *Yes* for Stimulus A is likely to disagree with the statements F and G. It might also have another group of subjects agreed with the statements expressing neutral attitude towards cats.

The family of Rasch models is widely used to analyze cumulative models, and the discrepancy between the empirical data patterns and the ideal Guttman patterns can be used for evaluating model-data fit. The unfolding models are developed to analyze unfolding data by assuming the underlying scale for the stimuli is unfolded and noncumulative response process.

Table 1 Cumulative and unfolding data

Type of Scale	Person	Stimulus							Stimulus Examples
		A	B	C	D	E	F	G	
Cumulative	**1**	1	0	0	0	0	0	0	A. 3 + 6 = ? B. 10 + 20 = ? C. 14 + 25 − 7 = ? D. 11 + 8 * 2 = ? E. 20 + 25 * 12 = ? F. (34 + 14 * 9)/8 = ? G. (53 – 11 * 4)/5 = ?
	2	1	1	0	0	0	0	0	
	3	1	1	1	0	0	0	0	
	4	1	1	1	1	0	0	0	
	5	1	1	1	1	1	0	0	
	6	1	1	1	1	1	1	0	
Unfolding	**1**	1	1	0	0	0	0	0	A. I hate cats. (Y/N) B. Cats are annoying. (Y/N) C. Cats are lazy. (Y/N) D. Cats are quiet. (Y/N) E. Cats like being clean. (Y/N) F. Cats are cute. (Y/N) G. I want to raise cats. (Y/N)
	2	1	1	1	0	0	0	0	
	3	0	1	1	1	1	0	0	
	4	0	0	0	1	1	1	1	
	5	0	0	0	0	1	1	1	
	6	0	0	0	0	0	0	1	

Within the context of accuracy studies for writing assessment, there are different ways of calculating accuracy data. One method put forward by Engelhard (1996) is to define accuracy ratings, A_{ni}, as

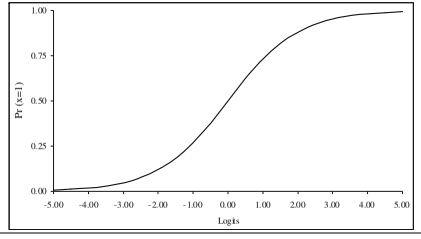
$$A_{ni} = \max\{|R_{ni} - B_i|\} - |R_{ni} - B_i|,$$

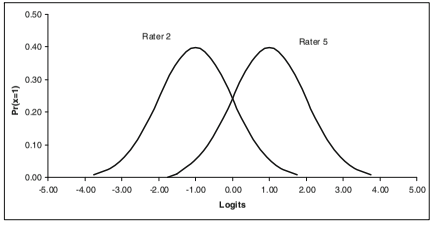
where R_{ni} is the observed rating of Rater n on Essay i, and B_i is the criterion rating on Essay i. In this way, all the possible values of accuracy ratings are in the positive direction. Using this approach, we obtained accuracy data. The hypothetical unfolding accuracy data in Table 1 within this context can be interpreted as dichotomous accuracy data that 0 represents inaccurate (i.e., does not match the criterion rating) and 1 represents accurate (i.e., match the criterion rating).

In order to illustrate the differences between a dichotomous Rasch model and an unfolding model, we fit a dichotomous Rasch model to the hypothetical unfolding accuracy data in Table 1 using the FACETS computer program (Linacre 2015). We also used an unfolding model (HCM) to estimate the parameters for the unfolding data using the RateFOLD computer program (Luo and Andrich 2003). There are several differences between two models shown in Table 2. For example, Raters 2 and 5 have the same location estimates (0.29 logits) with exactly the same item response function curves based on the Rasch model. When an unfolding model is used, then the location estimates for Raters 2 and 5 are −4.06, and 4.21, respectively. For the Rasch model, the probability response functions for Rater 2 and 5 are the same because they have the same accuracy measure. From the item response function curves obtained with the unfolding model, we can see that Raters 2 and 5 actually have different response functions. In other words, they have higher

Table 2 Analyses of hypothetical Unfolding Data

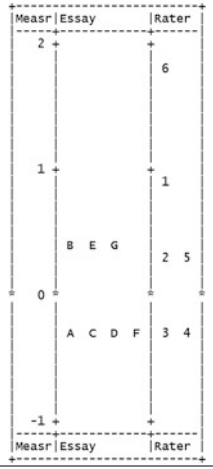
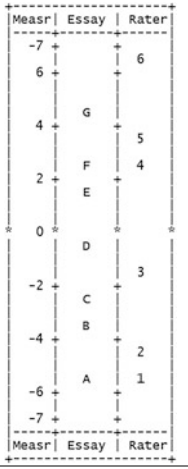
Rater	A	B	C	D	E	F	G	AR	RM	UM
1	1	1	0	0	0	0	0	28.6	0.94	−5.65
2	1	1	1	0	0	0	0	42.9	0.29	−4.06
3	0	1	1	1	1	0	0	57.1	−0.30	−0.86
4	0	0	0	1	1	1	1	57.1	−0.30	2.96
5	0	0	0	0	1	1	1	42.9	0.29	4.21
6	0	0	0	0	0	0	1	14.3	1.83	7.30
AR	33.3	50.0	33.3	33.3	50.0	33.3	50.0	–	–	–
RM	−0.33	0.44	−0.33	−0.33	0.44	−0.33	0.44	–	–	–
UM	−5.46	−3.46	−2.00	0.62	1.89	3.39	5.03	–	–	–
Theoretical item response function curves for Rasch model (Rater 2 and 5): Curves are the same for the two raters						Theoretical item response function curves for unfolding model (Rater 2 and 5): Curves are different for the two raters				





(continued)

Table 2 (continued)

Wright Map for Rasch model	Wright Map for Unfolding model
	

Note AR Accuracy Rate (%), *RM* Rasch Measures, *UM* Unfolding Measures

probabilities of rating accurately on different sets of essays. Therefore, unfolding models can capture important information existing in the unfolding data that is not evident in the dichotomous Rasch model for accuracy ratings.

The Hyperbolic Cosine Unfolding Model

For the Rasch model with ordered response categories, the category response curves for each essay along the scale of all the raters can indicate which raters have higher probabilities to score accurately on each essay and which raters have higher probabilities to score inaccurately (Fig. 1-Panel A). The category response curves for each rater along the scale of all the essays can indicate which essays that a rater has higher probabilities of rating accurately, and which essays that she has higher probabilities to score inaccurately (Fig. 1-Panel B). The duality property of IRT allows us to find the most useful information for our research purposes. In this study, we would like to evaluate rater accuracy towards student writing assessments. By having the raters as the focus, the threshold parameter shown in the category response curve is associated with each rater.

The HCM is derived from the Rasch model for ordered response categories (Andersen 1977; Andrich 1978, 1979; Wright and Master 1982) by assuming the distance between the adjacent categories are equal. When there are only three categories with equal distance from the rater location to each thresholds, the model can be expressed as

$$P(x=0) = \frac{1}{1 + \exp(\gamma_i + \delta_n - \theta_i) + \exp[2(\delta_n - \theta_i)]}, \tag{1}$$

$$P(x=1) = \frac{\exp(\gamma_i + \delta_n - \theta_i)}{1 + \exp(\gamma_i + \delta_n - \theta_i) + \exp[2(\delta_n - \theta_i)]}, \tag{2}$$

$$P(x=2) = \frac{\exp[2(\delta_n - \theta_i)]}{1 + \exp(\gamma_i + \delta_n - \theta_i) + \exp[2(\delta_n - \theta_i)]}, \tag{3}$$

where

δ_n represents the location of Essay n,

θ_i represents the location of Rater i,

γ_i represents the distance from Rater i's location to each of the thresholds.

The basic idea of the derivation is to fold the model in order to match the data which are assumed to be unfolded already. A brief illustration of the derivation is shown below based on the work of Andrich and Luo (1993).

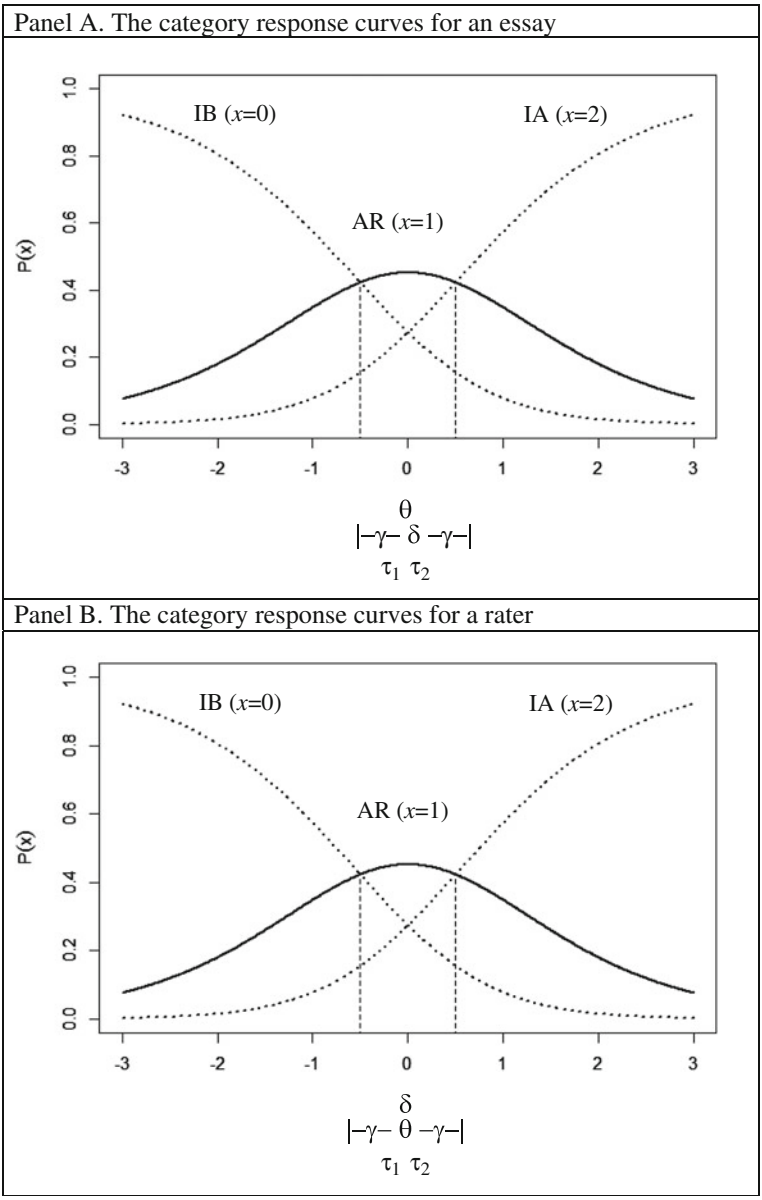


Fig. 1 Category response curves with rater focus and essay focus (*Note IB* Inaccurate Below, *IA* Inaccurate Above, *AR* Accurate Response)

Step 1: The sum of the probabilities of $x = 0$ and $x = 2$ (Eqs. 1 and 3) represents the probability of a single manifest inaccurate response which is scored as 0:

$$P(x=1) = \frac{\exp(\gamma_i + \delta_n - \theta_i)}{1 + \exp(\gamma_i + \delta_n - \theta_i) + \exp[2(\delta_n - \theta_i)]}, \tag{4}$$

$$P(x=0) = 1 - P(x=1) = \frac{1 + \exp[2(\delta_n - \theta_i)]}{1 + \exp(\gamma_i + \delta_n - \theta_i) + \exp[2(\delta_n - \theta_i)]}. \tag{5}$$

Step 2: Multiply the term $\exp(-\delta_n + \theta_i)$ for both numerator and denominator:

$$P(x=1) = \frac{\exp(\gamma_i)}{\exp(-\delta_n + \theta_i) + \exp(\gamma_i) + \exp(\delta_n - \theta_i)}, \tag{6}$$

$$P(x=0) = \frac{\exp(-\delta_n + \theta_i) + \exp(\delta_n - \theta_i)}{\exp(-\delta_n + \theta_i) + \exp(\gamma_i) + \exp(\delta_n - \theta_i)}. \tag{7}$$

Step 3: Based on the relationship between the exponential function and the hyperbolic cosine function, we have $[\exp(-a) + \exp(a)]/2 = \cosh(a)$ with $a = \delta_n - \theta_i$. Therefore, the HCM with a unit parameter γ_i becomes

$$P(x=1) = \frac{\exp(\gamma_i)}{\exp(\gamma_i) + 2\cosh(\delta_n - \theta_i)}, \tag{8}$$

$$P(x=0) = \frac{2\cosh(\delta_n - \theta_i)}{\exp(\gamma_i) + 2\cosh(\delta_n - \theta_i)}. \tag{9}$$

Step 4: Luo (1998) introduced a parameter ρ_i which is called the latitude of acceptance within the context of attitude measurement. We label it as the zone of accuracy for raters within rater-mediated assessments. The category response curves have two intersections when $p = 0.50$. So the distance between δ_n and θ_i satisfies $|\delta_n - \theta_i| = \rho_i$. The parameter ρ_i can be calculated by setting one of the probability functions equal to 0.50:

$$P(x=1) = \frac{\exp(\gamma_i)}{\exp(\gamma_i) + 2\cosh(\delta_n - \theta_i)} = \frac{\exp(\gamma_i)}{\exp(\gamma_i) + 2\cosh(\rho_i)} = \frac{1}{2}, \tag{10}$$

therefore

$$\exp(\gamma_i) = 2\cosh(\rho_i). \tag{11}$$

Step 5: Substitute Eq. (11) into Eqs. (8) and (9), the reparameterized HCM is formed as:

$$P(x=1)=\frac{\cosh(\rho_i)}{\cosh(\rho_i)+2\cosh(\delta_n-\theta_i)}, \tag{12}$$

$$P(x=0)=\frac{\cosh(\delta_n-\theta_i)}{\cosh(\rho_i)+\cosh(\delta_n-\theta_i)}. \tag{13}$$

The estimation method in use is the joint maximum likelihood estimation method with Newton-Raphson iteration algorithm. With the probability functions, one can get the information function for HCM based on Samejima's definition (1969). The general form of information function for a class of unfolding models was introduced by Luo and Andrich (2005):

$$I_{ni}=f(\pi_{ni})\Delta^2(\delta_n-\theta_i), \tag{14}$$

where

$$\pi_{ni}=P\{x_{ni}=1|\delta_n,\theta_i,\rho_i\}=\frac{\Psi(\rho_i)}{\Psi(\rho_i)+\Psi(\delta_n-\theta_i)}, \quad \text{and}$$
$$\Delta(t)=\frac{\partial\log\Psi(t)}{\partial t}=\frac{\partial\Psi(t)/\partial t}{\Psi(t)}.$$

For HCM,

$$\Delta(t)=\frac{\partial\cosh(\delta_n-\theta_i)/\partial t}{\cosh(\delta_n-\theta_i)}=\frac{\sinh(\delta_n-\theta_i)}{\cosh(\delta_n-\theta_i)}=\tanh(\delta_n-\theta_i),$$

therefore, the information function for HCM is

$$I=I_{ni}=P_{ni}(1-P_{ni})\tanh^2(\delta_n-\theta_i). \tag{15}$$

The goodness of fit can be evaluated in two ways: (a) the Pearson χ^2 statistic that not being significant indicates an acceptable overall fit of the model to the data and (b) a likelihood ratio χ^2 test that examine if the unit parameter ρ_i (i.e., zone of accuracy) is equal across all raters.

Using HCM for a Statewide Writing Assessment

The data analyzed in this section are based on the essays obtained from 8th grade students (N = 50) rated by randomly selected operational raters (N = 20) from a large-scale statewide writing assessment (Gyagenda and Engelhard 2010). The criterion scores were resolved by a validity panel of experts. Only the domain of

Style is selected, and dichotomized accuracy ratings (0-inaccurate, 1-accurate) are analyzed in the RateFOLD computer program (Luo and Andrich 2003).

First of all, the likelihood ratio χ^2 test for equal units is significant indicating that the zone of accuracy is different across raters. Therefore, the estimates and results based on the variant units are the optimal option, and they are summarized here. The overall test of fit based on the Pearson χ^2 statistic is not significant indicating an acceptable model-data fit.

From the Wright map, we can clearly see that raters are separated into three groups (Fig. 2). A follow-up qualitative study would be necessary to investigate similarities of the raters within each group, and to explore differences between the groups in terms of rater cognitive processes during scoring and perceptions towards essay characteristics. As indicated in Table 3, Rater 8 has the lowest estimate at −4.72 logits with a unit of 3.00 logits, while Rater 16 has the highest estimate at 3.75 logits with a unit of 2.54 logits. For Rater 1 and 18, the Chi-square tests for their location estimates are statistically significant, which indicates model-data misfit for these two raters.

The probability response function curves of unfolding models look like unfolded probability curves of Rasch models (Fig. 3). The location and the zone of accuracy of each rater are our focus. The location shows which essay(s) a rater has the highest probability to score accurately. The zone of accuracy indicates the range of essays that a rater has higher than 0.50 probabilities to score accurately. For example, Rater 19 is located at −4.46 logits and Rater 1 is located at 3.11 logits.

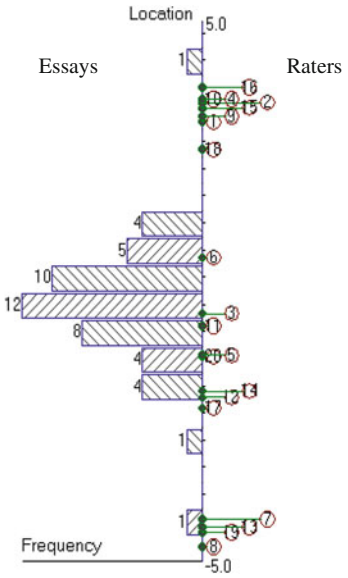


Fig. 2 Wright map for the HCM model

Table 3 Summary of rater measures with HCM

Raters	Accuracy Rates (%)	Location	Std. err	Unit	Chisq
1	58.0	3.11	0.16	3.68	10.03*
2	36.0	3.47	0.16	2.94	2.90
3	42.0	−0.41	0.24	0.38	4.46
4	30.0	3.54	0.17	2.69	4.32
5	52.0	−1.18	0.19	1.43	4.72
6	44.0	0.61	0.23	0.78	2.52
7	24.0	−4.21	0.19	2.64	1.16
8	22.0	−4.72	0.19	3.00	1.80
9	42.0	3.22	0.16	3.00	2.99
10	32.0	3.54	0.17	2.80	1.07
11	40.0	−.65	0.23	0.28	1.88
12	24.0	−1.97	0.20	0.07	2.86
13	28.0	−4.36	0.18	3.04	6.97
14	38.0	−1.85	0.18	1.21	2.24
15	34.0	3.37	0.16	2.74	0.62
16	24.0	3.75	0.18	2.54	3.42
17	22.0	−2.16	0.21	0.03	4.78
18	12.0	2.60	0.22	0.02	14.48*
19	28.0	−4.46	0.18	3.13	2.81
20	32.0	−1.22	0.21	0.02	1.82

*$p < 0.05$

These two raters are likely to be accurate in scoring different types of the essays. For a certain essay (e.g., an essay location with 2.0 logits), Rater 19 tends to score inaccurate above the criterion score and Rater 1 tends to be accurate on this essay scoring. When comparing the probability curves of Rater 19 and Rater 20, the zone of accuracy differs a lot. Rater 19 has a relatively large zone of accuracy with 3.13 logits, and this means that this rater tends to be accurate across more essays. Rater 20, on the other hand, has a smaller zone of accuracy indicating that this rater is accurate on a limited range of essays.

The information function shows the precision of the estimates that we have at each location point (Fig. 3). The rater information equation has a zero at the essay whose location equals to the rater location, and it has a maximum at the essay which has a distance from the rater location equals to its zone of accuracy. The theoretical maximum value of the information function for HCM is 0.25, because the range for math function $tanh(x)$ is between −1 and 1. Similarly, the information function curves of unfolding models seem like unfolded information curves of Rasch models.

The expected curve displays, if the observed rating scores of a rater match the expected rating scores from the model (Fig. 4). In a sense, it provides the

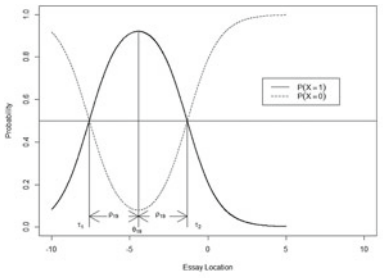
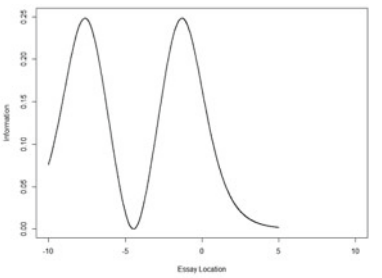
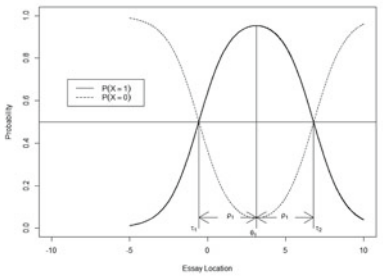
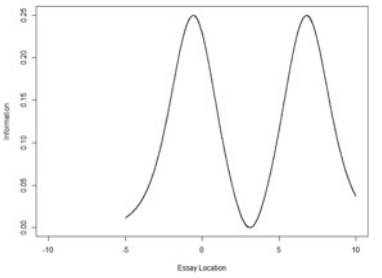
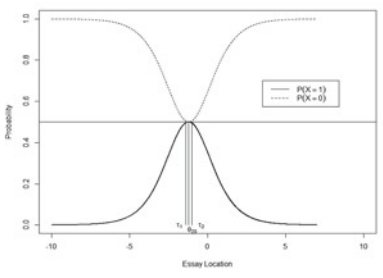
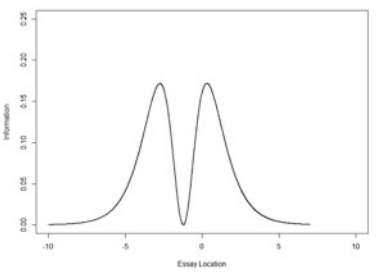
Rater	Probability function curve	Information function curve
19		
	θ_{19}= -4.46, ρ_{19}=3.13; accurate on essays located between τ_1 and τ_2	
1		
	θ_1= 3.11, ρ_1=3.68; accurate on essays located between τ_1 and τ_2	
20		
	θ_{20}= -1.22, ρ_{20}=.02; accurate on essays located between τ_1 and τ_2	

Fig. 3 Probability functions of selected raters

information for evaluating rater misfit together with the Chi-squared statistics. Based on the Chi-squared statistic, we know that Rater 1 exhibits misfit. From the expected curve of HCM, the distances between the observed values and the expected curve are quite large. For Rater 4, the observed values are closer to the expected curve compared with Rater 1, so it shows better model fit for Rater 4. Rater 15 has the best fit among these three raters, since the values fall close to the expected curve and the Chi-squared statistic is the smallest.

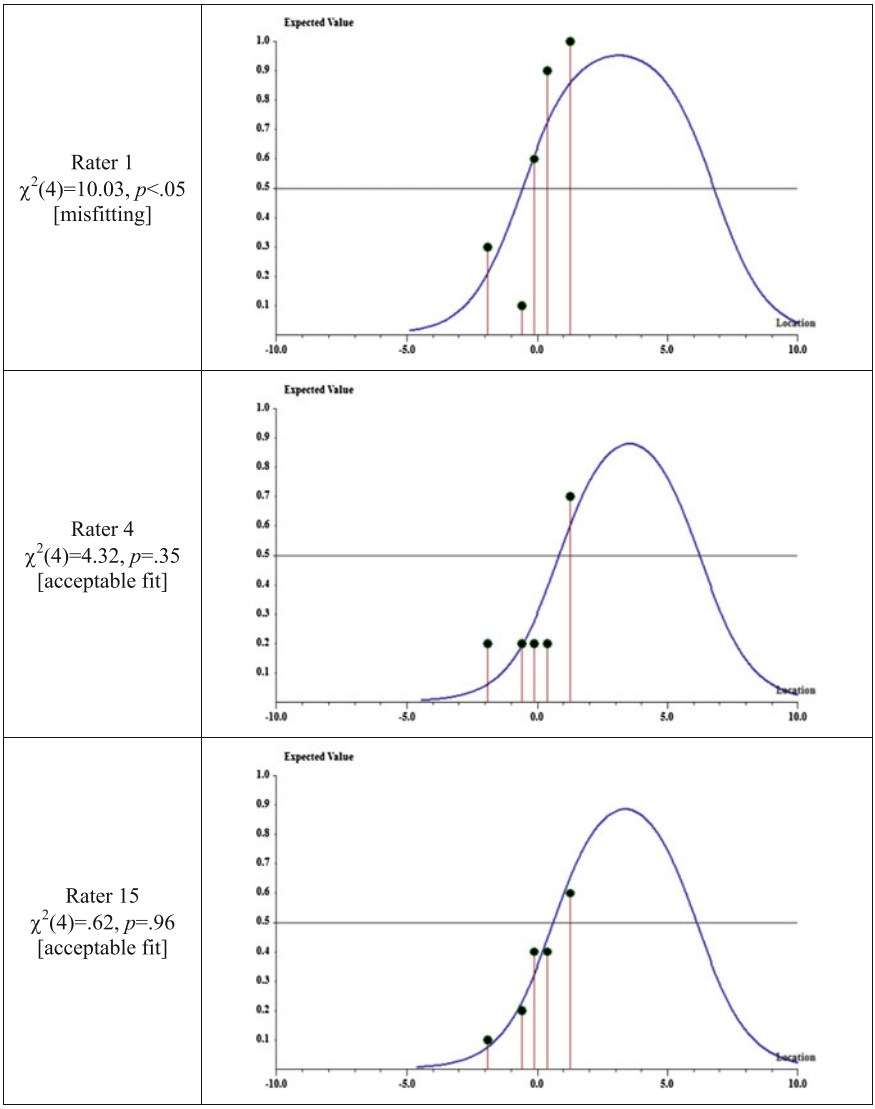


Fig. 4 Rater expected curves

Discussion

The HCM is one type of unfolding models, and it is introduced and illustrated as a tool to evaluate rater accuracy in this study. Accuracy is defined as a latent trait of raters. Unfolding models provide the opportunity to separate inaccuracy into inaccurate below and inaccurate above categories. Within the context of writing assessment,

raters assign ratings for the essays based on the rubrics. Therefore, it is similar to a preference study investigating if a rater prefers a certain feature of the essay to other features. No matter how much we want raters to consistently and precisely use the rubrics, the assigning of ratings will still have individual differences because of different human judgments. For example, raters differ in their prior knowledge and experience as well as the environmental factors related to the features of essays being rated. We view the current study as an exploratory analysis with a focus on identifying an underlying scale based on the empirical data. Results of this study indicate several possibilities for evaluating ratings of writing assessments which are viewed as unfolding data. Overall model-data fit was acceptable indicated by the Pearson Chi-squared statistic. However, the nature of the data depends on the nature of the task, and therefore one can decide which type of model (cumulative or unfolding) is appropriate to apply (Andrich 1988). Even though Andrich (1988) provided another way to check the underlying scale which is to compare the observed response patter with the theoretical response pattern (e.g., Guttman pattern or parallelogram), it is difficult to figure out the empirical data pattern with large-scale statewide writing assessments. Therefore, we suggest future research including a qualitative study of rater cognition and judgments to investigate the substantive meaning of the underlying scale.

This study describes how to use the HCM, and to interpret the results within the context of a writing assessment. Information provided by HCM offers new indices for evaluating rater accuracy. For example, the rater accuracy locations and the zones of accuracy provide direct measures for evaluating rating quality and rater performance. These accuracy indices can be used to guide future studies of rater cognition and judgment, as well as suggest new ways to evaluate and facilitate rater training in operational assessments.

Acknowledgments Pearson provided support for this research. Researchers supported by Pearson are encouraged to freely express their professional judgment. Therefore, the points of view or opinions stated in Pearson supported research do not necessarily represent official Pearson position or policy. We would like to thank Professors David Andrich and James Roberts for helpful comments and discussions of unfolding models.

References

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69–81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 357–374.

Andrich, D. (1979). A model for contingency tables having an ordered response classification. *Biometrics*, 403–415.

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12*(1), 33–51.

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement, 19*(3), 269–290.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17*(3), 253–276.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Coombs, C. H., & Avrunin, C. S. (1977). Single-peaked functions and the theory of preference. *Psychological Review, 84*(2), 216–230.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*, 177–193.

Davison, M. L. (1977). On A metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika, 42*, 523–548.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56–70.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Gyagenda, I. S., & Engelhard, G. (2010). Rater, domain, and gender influences on the assessed quality of student writing. In M. Garner, G. Engelhard, M. Wilson & W. Fisher (Eds.). *Advances in Rasch measurement* (pp. 398–429, Vol. 1). JAM press.

Jones, A. B. (2007). *Examining rater accuracy within the content of a high-stakes writing assessment (Unpublished doctoral dissertation)*. Atlanta, GA: Emory University.

Linacre, J. M. (2015). *Facets computer program for many-facet Rasch measurement*, version 3.71.4. Beaverton, Oregon: Winsteps.com.

Luo, G. (1998). A general formulation for unidimensional unfolding and pairwise preference models: making explicit the latitude of acceptance. *Journal of Mathematical Psychology, 42* (4), 400–417.

Luo, G. (2001). A class of probabilistic unfolding models for polytomous responses. *Journal of Mathematical Psychology, 45*(2), 224–248.

Luo, G., & Andrich, D. (2003). *RateFOLD computer program*. Social Measurement Laboratory: School of Education, Murdoch University. Western Australia.

Luo, G., & Andrich, D. (2005). Information functions for the general dichotomous unfolding model. In S. Alagumalai, et al. (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 309–328). Netherlands: Springer.

Murphy, K. R., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Poole, K. T. (2005). *Spatial models of parliamentary voting*. New York: Cambridge University Press.

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*, 231–255.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26* (2), 192–207.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. No: Psychometrika monograph supplement. 17.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497–506.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 278–286.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude: A psychophysical method and some experiments for measuring attitude toward the church*. Chicago: The University of Chicago Press.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31–37.

Wright, B.D., & Masters, G.N. (1982). Rating scale analysis: Rasch Measurement. Chicago: MESA Press.

Analyses of Testlet Data

Wen-Chung Wang and Kuan-Yu Jin

A testlet is a set of items that are linked by a common stimulus (e.g., a passage or a figure). The testlet design is widely used in educational and psychological tests. In mathematic tests or reading comprehension tests, for example, students may be required to respond to a few multiple-choice (MC) items based on the information provided in a figure, table, or reading passage. In addition to this, testlets have various other formats. A cloze test, in which some words in a passage are removed and are to be filled in, can be treated as a testlet (Baghaei 2008; B. Zhang 2010). Likewise, a C-test, which is a variant of cloze tests (Klein-Braley 1997; Schroeders et al. 2014), is another format of testlets. Paragraph-reorganizing items can be viewed as a testlet as well, in which disordered sentences or paragraphs are to be reorganized to form a meaningful passage (Min and He 2014). A common feature in these testlet formats is that all items in a testlet are linked, which may cause dependence or chain effects among items.

Within the framework of item response theory (IRT), practitioners often fit the 1-, 2-, or 3-parameter logistic model to dichotomous items, which is defined as follows:

$$P_{ni1} = c_i + (1 - c_i) \times \frac{\exp[a_i(\theta_n - b_i)]}{1 + \exp[a_i(\theta_n - b_i)]}, \quad (1)$$

P_{ni1} is the probability of scoring 1 (being correct) on item i for person n; θ_n is the latent trait of person n and is often assumed to follow a normal distribution; a_i is the slope (discrimination) parameter, b_i is the location (difficulty) parameter, and c_i is

W.-C. Wang (✉)
Department of Psychological Studies, Education University of Hong Kong,
10 Lo Ping Road, Tai Po, New Territories, Hong Kong
e-mail: wcwang@eduhk.hk

K.-Y. Jin
Education University of Hong Kong, Tai Po, Hong Kong

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_13

the asymptotic (pseudo-guessing) parameter of item i. Equation 1 is referred to as the 3-parmeter logistic model (3PLM). If $c_i = 0$ for every item, then Eq. 1 reduces to the 2-parameter logistic model (2PLM) (Birnbaum 1968). If $c_i = 0$ and $a_i = 1$ for every item, then Eq. 1 reduces to the 1-parmaeter logistic model (1PLM), which is also called the Rasch model (Rasch 1960).

Local item independence is a major assumption in IRT models or the family of Rasch models. That is, the item score residuals should be independent between items, after an IRT model is fit. If the residuals are dependent, the corresponding items are deemed as exhibiting local item dependence (LID). Items within the same testlet are likely to exhibit LID because they do not stand-alone; instead, they are linked by the same stimulus. It is commonly found that fitting a standard-IRT model (such as the 1PLM, 2PLM, and 3PLM) to testlet-based dichotomous items (referred to as the standard-IRT approach hereafter) often results in LID and biased parameter estimates (Sireci et al. 1991; Wainer et al. 2000).

In addition to the naive standard-IRT approach to testlet data, historically, one can transform a testlet with several dichotomous items into a super (polytomous) item and then fit a polytomous IRT model (Wainer and Kiely 1987), such as the (generalized) partial credit model, the graded response model, and the rating scale model (Andrich 1978; Masters 1982; Muraki 1992; Samejima 1969). This is referred to as the polytomous-item approach. For example, in a testlet consisting of five dichotomous items, the possible sum of the scores will be 0, 1, 2, 3, 4, and 5. This testlet is then considered as a polytomous-item with six ordered categories to which a polytomous IRT model can be fit, such as the generalized partial credit model (Muraki 1992), which is defined as follows:

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_i(\theta_n - b_{ij}), \tag{2}$$

where P_{nij} and $P_{ni(j-1)}$ are the probability of scoring j and $j - 1$ for person n on item i, respectively; θ_n is the latent trait of person n and is often assumed to follow a normal distribution; a_i and b_i are the discrimination and difficulty parameters of item i, respectively. When $a_i = 1$ for all polytomous-items, Eq. 2 reduces to the partial credit model (PCM; Masters 1982).

Testlet-based items themselves can be polytomous. If so, in the standard-IRT approach, the GPCM or the PCM can be fit, and in the polytomous-item approach, a super item needs to be created for each testlet. For example, if a testlet consists of two dichotomous items and one four-category item, the super item will consist of seven ordered categories (the sum of the scores of this super item ranges from 0 to 6). Then, in the polytomous-item approach, the GPCM or the PCM can be fit to such a super item as usual.

The polytomous-item approach is easy to implement for practitioners who are familiar with IRT. Any kind of LID within a testlet is absorbed into the summed score (Yen 1993). However, individual items within a testlet become invisible and

their item parameters (e.g., the parameters a and b of the original items) cannot be obtained.

The polytomous-item approach has been applied to empirical data (Keller et al. 2003; Thissen et al. 1989; Wainer and Lewis 1990; B. Zhang 2010). For example, B. Zhang adopted the polytomous-item approach to testlets in the Examination for the Certificate of Proficiency in English, which is an English language proficiency test for adult nonnative speakers of English at the advanced level (English Language Institute 2006). The test measures English language proficiency in speaking, writing, listening, and reading, and consists of 35 independent (stand-alone) items and 15 testlet-based items in three long dialogs or paragraphs. The 20 items in the cloze subtest sharing the same stimulus were treated as a testlet. In addition, the reading subtest that asks examinees to read four paragraphs, each followed by five questions, was treated as four testlets. Keller et al. (2003) analyzed the Auditing Section of the November 1998 administration of the Uniform Certified Professional Accountants Exam, which contained 75 MC items, 3 objective-answer formats, and 2 essays. Each objective-answer format was treated as a testlet. The total score of all dichotomous items within an objective-answer format was treated as a polytomous-item and analyzed with the polytomous-item approach.

The third approach to testlet data is to add an additional random-effect variable to standard-IRT models (e.g., 1PLM, 2PLM, and 3PLM, respectively) to account for LID within a testlet. These resulting testlet response models form testlet response theory (Bradlow et al. 1999; Li et al. 2006; Wainer et al. 2000; Wang and Wilson 2005). For example, a random-effect variable γ can be added to the 3PLM, one variable for each testlet, as follows:

$$P_{ni1} = c_i + (1 - c_i) \times \frac{\exp[a_i(\theta_n - b_i + \gamma_{nd(i)})]}{1 + \exp[a_i(\theta_n - b_i + \gamma_{nd(i)})]}, \tag{3}$$

$$\gamma_{nd(i)} \sim N\left(0,\ \sigma_d^2\right), \tag{4}$$

where $\gamma_{nd(i)}$ represents the interaction between person n and item i within testlet d and is assumed to follow a normal distribution with a mean of zero and variance σ_d^2 and is independent of θ; others have been defined in Eq. (1). Equation 3 together with Eq. 4 is called the 3-parameter testlet response model (3PTM). If $c_i = 0$ for every item, it can be reduced to the 2-parameter testlet response model (2PTM), and if $c_i = 0$ and $a_i = 1$ for every item, it can be reduced to the 1-parameter testlet response model (1PTM) (Wainer et al. 2007; Wang and Wilson 2005). The parameter σ_d^2 depicts the testlet effect (the magnitude of LID) in testlet d: the larger the variance, the larger the testlet effect. When σ_d^2 is zero for all testlets, the testlet response models are simplified to their corresponding standard-IRT models.

When testlet-based items are polytomous, one can add an additional random-effect variable to standard polytomous IRT models (Wang and Wilson 2005) as follows:

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_i(\theta_n - b_{ij} + \gamma_{nd(i)}), \tag{5}$$

with the assumption made in Eq. (4). Each item can maintain its parameters (e.g., the parameters a, and b) in this "testlet" approach. However, the testlet effect may not be fully accounted for by the incorporation of a random-effect parameter. For example, the testlet approach does not account for chain effects between two adjacent items, as explained below. It is difficult to implement the testlet approach when there are many testlets (i.e., computation burden due to many random-effect variables), which is often the case in large-scale tests. Furthermore, practitioners may not be very familiar with the corresponding computer programs for testlet response models and how to interpret their outputs.

It is possible that items within a testlet have chain effects, which means that a correct or incorrect answer to an item may increase or decrease the chance of success on other items in the same testlet (Yen 1993). This is especially likely to occur in cloze tests or C-tests. To account for the chain effects among items in a testlet, item-bundle IRT models have been developed in which a fixed-effect parameter is incorporated to describe each of the possible response patterns (Hoskens and De Boeck 1997; Tuerlinckx and De Boeck 1999; Wilson and Adams 1995). An example of this would be a testlet with two dichotomous items: Let $P(s, t)$ be the joint probability of scoring s on item 1 and t on item 2. There are four response patterns for the two dichotomous items and their joint probabilities, denoted as $P(0, 0)$, $P(0, 1)$, $P(1, 0)$, and $P(1, 1)$. The four joint probabilities can be modeled as follows:

$$\log[P(1,0)/P(0,0)] = \theta_n - b_1, \tag{6}$$

$$\log[P(0,1)/P(0,0)] = \theta_n - b_2, \tag{7}$$

$$\log[P(1,1)/P(0,0)] = (\theta_n - b_1) + (\theta_n - b_2) - b_{12}, \tag{8}$$

where θ_n is the latent trait of person n, b_1 is the difficulty of item 1, b_2 is the difficulty of item 2, and b_{12} describes the chain effect (interaction) between items 1 and 2. A negative b_{12} increases the probability of scoring 1 on both items, which suggests that a correct answer on item 1 helps answer item 2 correctly. In contrast, a positive b_{12} decreases the probability of scoring 1 on both items, which suggests that a correct answer on item 1 hinders the answering of item 2 correctly (this may not be very likely to occur in practice). The more extreme the value of b_{12}, the larger the chain effect between items. If b_{12} is zero, then these two items will be locally independent when standard-IRT models are fit.

If a testlet consists of two dichotomous items, then there will be 2^2 response patterns, and $2^2 - 1$ location parameters (i.e., b_1, b_2, and b_{12}) will be needed to describe these response patterns. The number of location parameters increases exponentially when the number of items increases linearly (in contrast, the number

of location parameters in the polytomous-item approach increases only linearly). For example, if a test consists of 10 dichotomous items, then there will be 2^{10} response patterns, and 2^{10-1} location parameters will be needed. A large proportion of response patterns may have a very small or even zero frequency. Therefore, this "response-pattern" approach becomes difficult to manage when the number of dichotomous items within a testlet is large (for example, if there are more than five items).

The standard-IRT, the polytomous-item, and the testlet approaches to testlet data can yield very different results. Schroeders et al. (2014) analyzed 12 German C-tests with the three approaches: (a) the standard-IRT approach based on the 1PLM, (b) the testlet approach based on the 1PTM, and (c) the polytomous-item approach based on the PCM. When the standard-IRT approach was adopted, item residuals were still correlated, which suggests LID. Furthermore, the item parameter estimates obtained from the three approaches were very different, which was expected because they were on different scales, but the person measures obtained from the three approaches were almost perfectly correlated ($r > 0.99$). A simulation study was conducted to verify that ignoring testlet effects by fitting the 1PLM to the IPTM data tended to lead to overestimation of the test reliability. O. Zhang (2010) fitted the 1PLM, the 1PTM, and the PCM to the data that simulated data from the 1PTM. It was concluded that fitting the 1PLM tended to lead to overestimation of the test reliabilities and underestimation of the standard errors for person measures, whereas fitting the 1PTM and PCM yielded unbiased estimates for the test reliabilities and standard errors.

In summary, the standard-IRT approach, although very easy to implement, fails to account for LID; the polytomous-item approach is easy to implement at the expense of invisible items; the response-pattern approach can address the LID within a testlet thoroughly but is difficult to manage when there are many items in a testlet; the testlet approach preserves individual items but fails to account for chain effects and is difficult to implement for many practitioners. It should be noted that the item parameters in these approaches are not comparable because they are on different scales. In cases where the estimation of item parameters is the key concern (e.g., item parameters must be on the same scale when different items are to be linked to form an item bank), the same approach should be applied to all testlet datasets. On the other hand, if the estimation of person measures as well as test reliability, rather than the estimation of item parameters, is the key concern (e.g., ranking test-takers is the most important purpose in educational testing), perhaps the polytomous-item approach should be chosen.

In theory, fitting a true model to data will yield the best results than fitting wrong models. For example, fitting the 1PLM to data simulated from the 1PLM and fitting the 1PTM to data simulated from the 1PTM will yield the best results than fitting other models. Unfortunately, true models are never known in practice. It is thus very critical to evaluate the robustness of different approaches to testlet data, given that all these approaches are based on wrong models. For example, if data follow

the response-pattern models (i.e., chain effects among items within a testlet) but are analyzed with three wrong models (the 1PLM, 1PTM, and PCM), it is not clear which of the three models would yield the best results. We conducted a series of simulations to evaluate the standard-IRT, the polytomous-item, and the testlet approaches to testlet data.

Simulation Studies

Design

Two simulation studies were conducted. In both studies, there were a total of 2000 examinees who responded to 40 dichotomous items. In study 1, item responses were generated from the 1PTM, 2PTM, or 3PTM. The simulation conditions are summarized in Table 1. The standard-IRT approach (the 1PLM, 2PLM, or 3PLM), the polytomous-item approach (the PCM or GPCM), and the testlet approach (the true models) were fit to the data. Specifically, when data were simulated from 1PTM, the standard-IRT approach would be based on the 1PLM, the polytomous-item approach would be based on the PCM, and the testlet approach would be based on the 1PTM. When data were simulated from 2PTM, the standard-IRT approach would be based on the 2PLM, the polytomous-item approach would be based on the GPCM, and the testlet approach would be based on the 2PTM. When data were simulated from 3PTM, the standard-IRT approach would be based on the 3PLM, the polytomous-item approach would be based on the GPCM, and the testlet approach would be based on the 3PTM.

In study 2, the 40 dichotomous items belonged to eight testlets, and each testlet had five dichotomous items. The data were simulated from the response-pattern approach as follows. Let $P(0, 0, 0, 0, 0)$ be the probability of scoring 0 on all of the five items. The chain effect between two adjacent items in the same testlet was modeled as follows:

Scoring 1 on two adjacent items:

$$\log[P(1,1,0,0,0)/P(0,0,0,0,0)] = (\theta_n - b_1) + (\theta_n - b_2) - \eta, \qquad (9)$$

Table 1 Six conditions in simulation study 1

	Condition					
	Nil	I	II	III	IV	V
No. of testlets	8	2	4	6	8	8
No. of independent items	0	30	20	10	0	0
Testlet variance	0	0.64	0.64	0.64	0.64	1.44

$$\log[P(0,1,1,0,0)/P(0,0,0,0,0)] = (\theta_n - b_2) + (\theta_n - b_3) - \eta, \tag{10}$$

$$\vdots$$

$$\log[P(0,0,0,1,1)/P(0,0,0,0,0)] = (\theta_n - b_4) + (\theta_n - b_5) - \eta. \tag{11}$$

Scoring 1 on three adjacent items:

$$\log[P(1,1,1,0,0)/P(0,0,0,0,0)] = (\theta_n - b_1) + (\theta_n - b_2) + (\theta_n - b_3) - 2\eta, \tag{12}$$

$$\vdots$$

$$\log[P(0,0,1,1,1)/P(0,0,0,0,0)] = (\theta_n - b_3) + (\theta_n - b_4) + (\theta_n - b_5) - 2\eta. \tag{13}$$

Scoring 1 on four adjacent items:

$$\log[P(1,1,1,1,0)/P(0,0,0,0,0)] = (\theta_n - b_1) + (\theta_n - b_2) + (\theta_n - b_3) + (\theta_n - b_4) - 3\eta, \tag{14}$$

$$\vdots$$

$$\log[P(0,1,1,1,1)/P(0,0,0,0,0)] = (\theta_n - b_2) + (\theta_n - b_3) + (\theta_n - b_4) + (\theta_n - b_5) - 3\eta. \tag{15}$$

Scoring 1 on five adjacent items:

$$\log[P(1,1,1,1,1)/P(0,0,0,0,0)] = (\theta_n - b_1) + (\theta_n - b_2) + (\theta_n - b_3) + (\theta_n - b_4) + (\theta_n - b_5) - 4\eta. \tag{16}$$

The log-odds for a response-pattern with a score of 1 for nonadjacent items did not consist of η; for example,

$$\log[P(1,0,1,0,1)/P(0,0,0,0,0)] = (\theta_n - b_1) + (\theta_n - b_3) + (\theta_n - b_5). \tag{17}$$

The formulation of the chain effects in Eqs. 9–17 is referred to as the 1-parameter response-pattern model (1PPM). In theory, there are 32 response patterns for the five dichotomous items so that altogether there can be 31 location parameters. That is, in addition to the five item parameters, b_1 to b_5, there can be 26 additional parameters to account for the chain effects among these 32 response-patterns. However, in Eqs. 9–17, all the chain effects are assumed to be accounted for by one single parameter η, which is of course a very stringent assumption. If η was zero, there was no interaction between items, so these items can be treated as independent items and fitting standard-IRT models would be appropriate. Here, η was set at −2, −1, 1, or 2. After data were generated from the 1PPM, the standard-IRT, the polytomous-item, and the testlet approaches were fit to the simulated data. For comparison, the true model was also fit.

Analysis

The freeware WinBUGS (Spiegelhalter et al. 2007) was used to estimate the parameters, for which the Markov chain Monte Carlo methods were implemented for parameter estimation. A total of ten replications were made in each study, because each replication required several hours of computing time. Because the person measures in the standard-IRT, the polytomous-item, the testlet, and the response-pattern approaches were not on the same scale, and the rank orders of person measures were commonly used to make decisions (e.g., college admission), we computed the mean absolute rank order change (*MARC*) in person measures across replications to compare the effectiveness of these approaches:

$$MARC = \sum_{s=1}^{S} \left(\left| \sum_{n=1}^{N} \hat{R}_n - R_n \right| / N \right) / S, \tag{18}$$

where $\hat{R}_n$ and R_n are the rank orders of the estimated and true measures of person n, respectively, N is the number of examinees, and S is the number of replications. The *MARC* obtained from fitting the true model to the data was treated as the gold standard, with which the *MARC* obtained from fitting the other models was compared. The closer the *MARC* was to the gold standard, the better the approach.

In addition to the *MARC*, we computed the test reliability for each approach, which was the square of the correlation between the estimated and true person measures. The test reliability determined by fitting the true model to the data was treated as the gold standard, with which the test reliability determined from fitting the other models was compared. The closer the test reliability was to the gold standard, the better the approach.

Results

Study 1. The *MARC* for the standard-IRT, polytomous-item, and testlet approaches is shown in Table 2. The *MARC* obtained from the testlet approach was treated as the gold standard because it was used to generate the data. The *MARC* for the standard-IRT and the polytomous-item approaches in Condition Nil (where the testlet variance was zero) was always the closest to the gold standard across different conditions, and the *MARC* showed greater differences with the gold standard as the testlet variance increased from Condition I to Condition V. When the data were generated from the 1PTM, both the standard-IRT approach (the 1PLM) and the polytomous-item approach (the PCM) yielded an *MARC* that was very close to the gold standard, with a difference that was no greater than 5. When the data were generated from the 2PTM or the 3PTM, the standard-IRT approach (the 2PLM or the 3PLM) yielded an *MARC* that was very close to the gold standard, with a difference that was no greater than 5. However, under the same setting, the polytomous-item approach (the GPCM) yielded an *MARC* that could be larger than the gold standard by 10.45. This large difference might be attributable to the presence of only a single discrimination parameter in the GPCM, but each individual item within a testlet was generated to have its own discrimination parameter.

The mean test reliabilities across ten replications obtained from the standard-IRT, the polytomous-item approaches, and the gold standard are shown in Table 3. Across conditions, the polytomous-item approach resulted in a small difference to the gold standard (−0.030 to 0.007). The standard-IRT approach resulted in a small difference to the gold standard (−0.009 to 0.010) only under Conditions Nil and I, in which the testlet effects were zero and small, respectively. As the testlet effect increased from Conditions II to V, the standard-IRT approach

Table 2 Mean absolute rank order changes across ten replications in simulation study 1

Approach (Model)	Condition					
	Nil	I	II	III	IV	V
Testlet (1PTM)[a]	153.27	162.59	177.87	189.20	207.62	241.52
Standard (1PLM)	0.26	3.10	4.91	3.47	−0.06	0.99
Polytomous-item (PCM)	0.28	3.03	5.02	3.40	−0.05	1.04
Testlet (2PTM)[a]	147.64	152.66	166.42	184.20	193.93	238.34
Standard (2PLM)	−0.19	1.54	3.98	4.33	0.80	2.45
Polytomous-item (GPCM)	7.53	0.38	0.78	2.43	3.88	1.88
Testlet (3PTM)[a]	187.98	187.96	191.73	204.84	209.58	246.74
Standard (3PLM)	0.17	1.27	2.37	2.67	0.87	0.37
Polytomous-item (GPCM)	10.45	1.85	1.38	3.51	6.82	3.86

Note [a]Gold standard; 1*PLM* 1-parameter logistic model; 2*PLM* 2-parameter logistic model; 3*PLM* 3-parameter logistic model; *PCM* partial credit model; *GPCM* generalized partial credit model; 1*PTM* 1-parameter testlet response model; 2*PTM* 2-parameter testlet response model; 3*PTM* 3-parameter testlet response model

Table 3 Mean test reliability estimates across ten replications in simulation study 1

Approach (Model)	Condition					
	Nil	I	II	III	IV	V
Testlet (1PTM)[a]	0.880	0.865	0.845	0.828	0.797	0.731
Standard (1PLM)	−0.004	0.010	0.023	0.036	0.055	0.115
Polytomous-item (PCM)	−0.004	0.004	0.007	0.004	−0.004	−0.007
Testlet (2PTM)[a]	0.885	0.880	0.861	0.830	0.815	0.734
Standard (2PLM)	−0.002	0.004	0.024	0.040	0.060	0.122
Polytomous-item (GPCM)	−0.014	−0.003	0.001	0.000	−0.004	−0.005
Testlet (3PTM)[a]	0.828	0.823	0.812	0.792	0.785	0.716
Standard (3PLM)	−0.009	0.001	0.020	0.030	0.044	0.091
Polytomous-item (GPCM)	−0.022	−0.008	−0.002	−0.011	−0.016	−0.030

Note [a]Gold standard; 1*PLM* 1-parameter logistic model; 2*PLM* 2-parameter logistic model; 3*PLM* 3-parameter logistic model; *PCM* partial credit model; *GPCM* generalized partial credit model; 1*PTM* 1-parameter testlet response model; 2*PTM* 2-parameter testlet response model; 3*PTM* 3-parameter testlet response model. The values for the standard and the polytomous-item approaches are differences in the mean test reliability between the corresponding approach and the gold standard

resulted in a larger difference to the gold standard (0.020 to 0.122). These findings indicate that the larger the testlet effect, the more serious the overestimation in test reliability using the standard-IRT approach.

In short, the polytomous-item approach appears to be feasible when the data are generated from the testlet approach, especially in the framework of Rasch models; however, the standard approach would result in substantial overestimation of the test reliability when the testlet effect is large.

Study 2. The *MARC* across ten replications for the standard-IRT, the polytomous-item, and the testlet approaches, as well as the response-pattern approaches is listed in Table 4, in which the response-pattern approach was treated

Table 4 Grand mean of the mean absolute rank order changes across ten replications in simulation study 2

Approach (Model)	$\eta = -2$	$\eta = -1$	$\eta = 1$	$\eta = 2$
Response-Pattern (1PPM)[a]	237.05	166.24	185.08	205.48
Standard (1PLM)	−0.63	−0.52	−0.15	−0.22
Polytomous-item (PCM)	0.43	−0.33	−0.29	−0.66
Testlet (1PTM)	−0.27	−0.30	−0.47	0.05

Note [a]Gold standard; 1*PPM* 1-parameter response-pattern model; *PCM* partial credit model; 1*PLM* 1-parameter logistic model; 1*PTM* 1-parameter testlet response model. The values for the standard, the polytomous-item, and the testlet approaches are differences in the mean absolute rank order changes between the corresponding approach and the gold standard

Table 5 Mean test reliability estimates across ten replications in simulation study 2

Approach (Model)	$\eta = -2$	$\eta = -1$	$\eta = 1$	$\eta = 2$
Response-pattern (1PPM)[a]	0.738	0.850	0.837	0.803
Standard (1PLM)	0.075	0.031	−0.057	−0.123
Polytomous-item (PCM)	0.001	−0.005	−0.002	−0.004
Testlet (1PTM)	0.032	0.005	−0.061	−0.129

Note [a]Gold standard; 1*PPM* 1-parameter response-pattern model; *PCM* partial credit model; 1*PLM* 1-parameter logistic model; 1*PTM* 1-parameter testlet response model. The values for the standard, the polytomous-item, and the testlet approaches are differences in the mean test reliabilities between the corresponding approach and the gold standard

as the gold standard because it was used to generate the data. It appeared that the three approaches yielded an *MARC* that was almost identical to the gold standard, with a difference of −0.66 to 0.43. According to the mean test reliability shown in Table 5, the polytomous-item approach yielded a test reliability almost identical to that of the gold standard, whereas the standard-IRT and the testlet approaches overestimated the test reliability when η was negative, but underestimated it when η was positive. The biased estimation of test reliability was more serious when the value of η was more extreme. For example, when $\eta = -2$, the standard-IRT and the testlet approaches resulted in overestimation of the test reliability by 0.075 and 0.032, respectively, whereas when $\eta = 2$, they resulted in underestimation of the test reliability by 0.123 and 0.129, respectively. In short, only the polytomous-item approach showed good performance when the data were generated using the response-pattern approach.

An Empirical Example

An English midterm exam in high schools was analyzed, which consisted of 77 items with three parts. The first part contained 59 independent MC items: students were required to select the best answer to each of the three- or four-option MC items. The second part contained three reading comprehension testlets: each testlet had a reading passage followed by four or five MC items, including childhood memory (testlet 1 with four MC items), job recruitment (testlet 2 with four MC items), and Mexicans' expression (testlet 3 with five MC items). All the MC items had four options. The third part contained five summary items: in each summary item, students were required to select the best answer from six options that expressed the main idea in the passage. The five summary items were treated as a

testlet. Altogether, the English test consisted of 59 independent MC items, three reading comprehension testlets, and one summary testlet.

A total of 1073 students from a province of China took the test, and their item responses were analyzed using the standard-IRT, the polytomous-item, the testlet, and the response-pattern approaches with WinBUGS. In the standard-IRT approach, the testlet-based items were treated as independent items and all items in the test were assumed to follow the 1PLM. In the polytomous-item approach, the MC items in a testlet were transformed into a polytomous-item and all items in the test were analyzed with the PCM (the 59 independent items were assumed to follow the 1PLM). In the testlet approach, the testlet-based items were assumed to follow the 1PTM, and the 59 independent items were assumed to follow the 1PLM. In the response-pattern approach, the chain effect between adjacent items in a testlet was modeled as the 1PPM (see Eqs. 9–17), and the 59 independent items were assumed to follow the 1PLM.

In the testlet approach, the variances in the additional random-effect parameters (σ_d^2) described the magnitudes of the testlet effects. In the response-pattern approach, the η parameters described the magnitudes of the chain effects. We were particularly interested in the rank orders of person measures and test reliability. Based on the findings in the two simulation studies, it was anticipated that (a) the person measures obtained from the four approaches would be highly correlated; (b) the rank orders of the person measures would not show a large difference among the four approaches; (c) if the testlet effects or the chain effects were substantial, the standard-IRT approach would yield a test reliability that was larger than that of the other three approaches.

Results

In the testlet approach, the estimates (and standard errors) for σ_d^2 of the four testlets were 0.19 (0.07), 1.51 (0.38), 0.08 (0.06), and 2.91 (0.50); the values indicate that the testlet effect was large in testlets 2 and 4. In the response-pattern approach, the estimates (and standard errors) for the η parameters of the four testlets were −0.10 (0.08), −1.95 (0.17), −0.10 (0.10), and −2.44 (0.16); the values indicate that a student who succeeded in a preceding item had a higher possibility of success on the subsequent items in testlets 2 and 4. Although these two approaches accounted for different types of LID, LID was considerably large in testlets 2 and 4. We computed the residuals between the observed sum scores and the expected sum scores for testlets 2 and 4, respectively, in the 1PLM, PCM, 1PTM, and 1PPM and drew box-plots (shown in Fig. 1). If the model had a good fit to the data, the

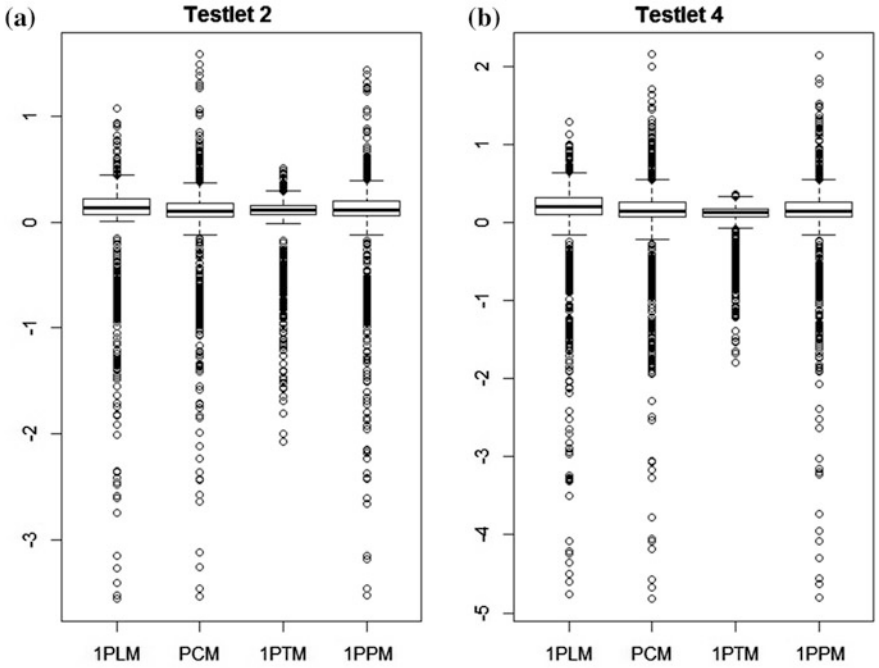


Fig. 1 Residuals for testlets 2 and 4 in the English exam for the four models. *Note* 1*PLM* 1-parameter logistic model; *PCM* partial credit model; 1*PTM* 1-parameter testlet model; 1*PPM* 1-parameter response-pattern model

residuals should be close to zero. It seemed that the 1PTM had the best fit among the four models.

The person measures derived from the four approaches, shown in Fig. 2, were almost perfectly correlated ($r > 0.996$). We used the rank orders obtained from the polytomous-item approach as the reference and computed the *MARC* for the standard-IRT, the testlet, and the response-pattern approaches, which was found to be 12.84, 14.57, and 13.07, respectively. An *MARC* between 12.84 and 14.57 in 1073 students can be considered as small.

The test reliabilities for the standard-IRT, the polytomous-item, the testlet, and the response-pattern approaches were 0.863, 0.855, 0.855, and 0.854, respectively. As expected, when there was substantial LID within testlets, the standard-IRT approach tended to result in overestimation of the test reliability, whereas the polytomous-item approach was not affected. In view of the overall performance and ease of usage, the polytomous-item approach appears very promising.

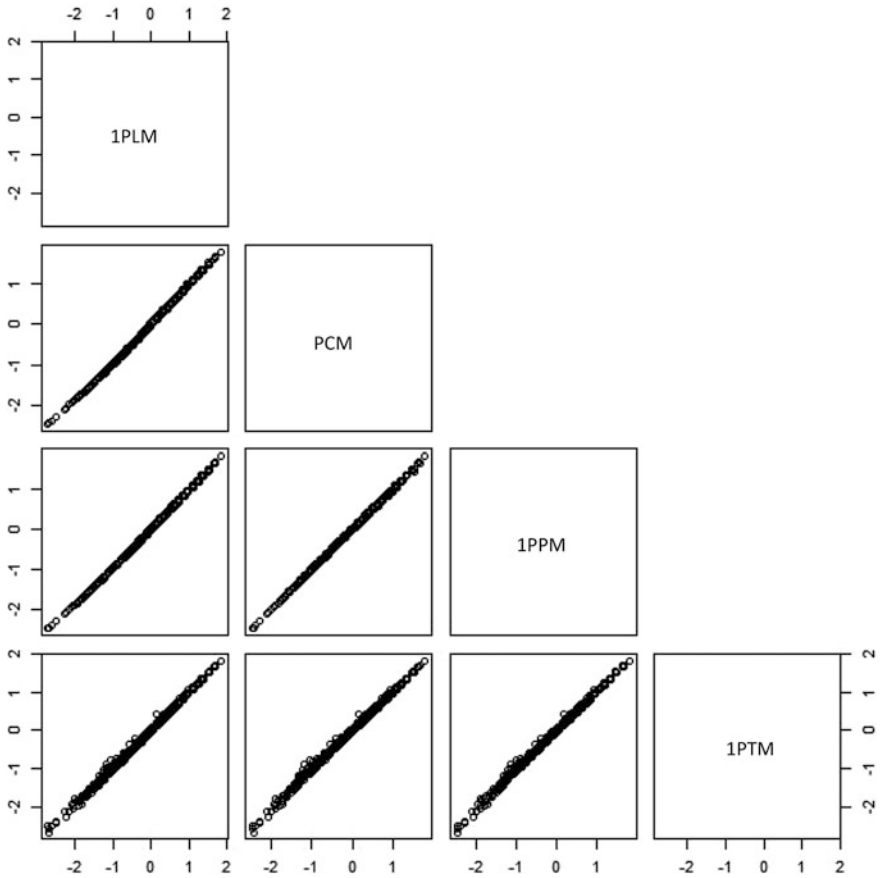


Fig. 2 Correlations of person measures in the English exam among the four models. *Note* 1*PLM* 1-parameter logistic model; *PCM* partial credit model; 1*PTM* 1-parameter testlet model; 1*PPM* 1-parameter response-pattern model

Conclusion and Discussion

There are pros and cons of the testlet and the polytomous-item approaches. Although the testlet approach keeps individual items within a testlet visible, many practitioners find it difficult to implement this approach. On the other hand, although the polytomous-item approach is easy to implement, it makes individual items within a testlet invisible. Sometimes, keeping individual items visible is crucial. For example, practitioners may want to know why an individual item is more difficult than another individual item: when testlet-based items have been analyzed with the testlet approach before and a few common testlet-based items are used in a new test to link the old and new tests on the same scale, the same testlet

approach has to be adopted. Sometimes, in other cases, ranking person measures is the key concern and individual items are not of interest at all, such as education admission. In such cases, the polytomous-item approach is a good choice. Of course, in test linking, if the polytomous-item approach has been adopted in an old test, the same approach has to be adopted again in a new test.

This study compared the standard-IRT, the polytomous-item, and the testlet approaches to testlet data for recovering person measures and their test reliabilities when data were generated from the testlet and the response-pattern approaches. Three major findings were obtained from the simulations. First, with the polytomous-item approach, the rank orders of person measures could be recovered almost as accurately as the gold standard (when the true model was fit to the simulated data), and the test reliability estimates obtained were almost the same as that of the gold standard. Second, with the standard-IRT approach, although the rank orders of person measures can be recovered very accurately, the test reliability is overestimated substantially when the data are generated from the testlet approach with large testlet effects, and the test reliability is substantially underestimated or overestimated when the data are generated using the response-pattern approach (according to the direction of the interaction term η). Third, similar to the standard-IRT approach, the testlet approach resulted in substantial underestimation or overestimation of the test reliability when the data were generated using the response-pattern approach (according to the direction of the interaction term η). Given that the polytomous-item approach performs fairly well and is easy to implement, and the testlet approach cannot fully capture local dependence due to the chain effects, the polytomous-item approach is recommended for testlet-based items, when the estimation of person measures and test reliability is the major concern.

References

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43*, 561–573.

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. 397–479). Reading, MA: Addison-Wesley.

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions, 21* (3), 1105–1106.

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

English Language Institute, University of Michigan. (2006). *Examination for the certificate of proficiency in English 2004–05 annual report*. Ann Arbor, MI: English Language Institute, University of Michigan.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods, 2*, 261–277.

Keller, L. A., Swaminathan, H., & Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in Education, 16*, 207–222.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14*, 47–84.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3–21.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*, 453–477.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monographs, 17*, 1–100.

Schroeders, U., Robitzsch, A., & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with C-tests. *Journal of Educational Measurement, 51*, 400–418.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2007). WinBUGS version 1.4.3: Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: a use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247–260.

Tuerlinckx, F., & De Boeck, P. (1999). Distinguishing constant and dimension-dependent interaction: A simulation study. *Applied Psychological Measurement, 23*, 299–307.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1–14.

Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & G. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Dordrecht: Springer Netherlands.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126–149.

Wilson, M., & Adams, R. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181–198.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27*, 119–140.

Zhang, O. (2010). *Polytomous IRT or Testlet Model: An evaluation of scoring models in small testlet size situations (Unpublished master's Thesis)*. Gainesville, FL: University of Florida.

From Standards to Rubrics: Comparing Full-Range to At-Level Applications of an Item-Level Scoring Rubric on an Oral Proficiency Assessment

Troy L. Cox and Randall S. Davies

Introduction

Standards-based proficiency frameworks have become an integral part of the educational assessment landscape. These frameworks take complex, multidimensional competencies and attempt to represent them as a numerical value on a vertical scale that can be used by students, teachers, testing organizations, school admissions officers, employers, and others that want some certification of the proficiency of examinees. Framing performance in this way allows stakeholders to communicate and compare results. With some frameworks, like the Common Core (National Governors Association Center for Best Practices & Council of Chief State School Officers 2010), the vertical axis of the scale is based on grade levels. With language proficiency, the vertical axis of the scales is based on major-level descriptors which define what an individual should be able to do if they are to be certified as being proficient in that language at a specific level [see for example, the Common European Framework of Reference (Verhelst et al. 2009) and the American Council on the Teaching of Foreign Languages (ACTFL 2012)].

Rubrics are an essential component of any framework (Bargainnier 2004; Tierney and Simon 2004). Practitioners attempting to assess any performance must use rubrics that align with the standards. Students wanting to improve their performance need to understand how their score relates to the standards. Test developers needing to create equivalent test forms must have a way to ensure those forms are based on the standards. To be useful, the relationship between the rubric and the

T.L. Cox (✉)
Center for Language Studies, Brigham Young University,
3086-C JFSB, Provo, UT 84602, USA
e-mail: troyc@byu.edu

R.S. Davies
Instructional Psychology and Technology, Brigham Young University,
150-L MCKB, Provo, UT 84602, USA

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_14

standard should be transparent and the way in which raters use the rubrics must be consistent and appropriate.

Often when assessing speaking ability, some type of oral proficiency interview (OPI) is used. For example, with an ACTFL OPI, trained interviewers prompt examinees to respond to a wide variety of tasks (Buck et al. 1989). Each speaking task is designed to target a specific level on the scale and is intended to provide the examinee with an opportunity to demonstrate his or her ability to speak at that level. When an individual responds well to a specific prompt it provides evidence the individual is able to speak at that level. The assessment is designed to push the limits of the examinee and determine when the individual's speaking ability breaks down. Using the evidence they have gathered, interviewers then rate the performance against the proficiency standards and assign a score based on the scale being used. The score provides an estimate of the general speaking ability of the examinee. Since interviews are expensive to administer, many testing companies are transitioning to computer-administered speaking tests which in turn affects the way the test can be rated.

Rating can be done at the test-level or at the item-level. When raters assess speaking competency at the test-level, the rater listens to the entire performance and, rating it holistically, determines the overall speaking ability or level of the examinee. However, if computer-aided assessments are to be used or equivalent forms of a test are needed, item-level assessments are required. When rating at the item-level, raters listen to an individual's response to a specific task rating each it against the rubric. The individual item scores are combined to determine the overall speaking level of the examinee.

One problem human raters have when rating at the item-level is how to apply the rubric. Raters might be inclined to rate the performance using the full-range of the rubric as they would when rating holistically at the test-level. However, individual task prompts are not designed to provide that type of evidence. For example, a task designed to elicit evidence that an examinee can speak at an Intermediate level on the ACTFL scale would not likely provide evidence that the individual can speak at a higher level (e.g., superior). The individual would need to be prompted with another task designed to elicit that type of evidence. This study examined two ways of applying a rubric at the item-level—one that was directly tied to the full-range of the proficiency scale and another that used a restricted-range of that same scale.

Research Questions

1. How reliable is the full-range proficiency-based rubric when used at the item level?
2. How reliable is the at-level proficiency-based rubric when used at the item level?
3. Which rubric (full-range or at-level) most closely aligned with the expert-predicted item difficulty (EID) of each prompt?

Methods

This paper examined two ways to implement a proficiency-based rubric (full-range and at-level) when rating at the item-level. The item difficulty statistics calculated from the two rubrics were compared to the intended item difficulty of the item writers. Finally, a comparison was made between examinee test scores that were scored using a full-range and an at-level restricted-range application of the rubric.

Study Participants

The subjects participating in this study were students enrolled at an intensive English Program affiliated with a large research university that were taking their exit exams during winter semester 2012. There were 201 students who spoke 18 different languages (see Table 1). They were in the school to improve their English to the point at which they could successfully attend university, where the language of instruction was English. With the ACTFL guidelines, they ranged in speaking ability from Novice to Advanced.

Data Collection Instrument

The data collection instrument used in this study was designed to assess speaking ability at proficiency levels 2 through 6 (see Appendix A). It was assumed that after one semester of instruction, all the examinees participating in this study would have some ability to speak English yet none would be considered the functional equivalent of highly educated native speakers.

To determine to what extent the prompts on the instrument aligned with their expected difficulty level, a panel of expert raters was consulted. The rating rubric

Table 1 Composition of subjects by language and gender

Native language	Gender		Total
	Female	Male	
Spanish	54	34	88
Korean	21	16	37
[a]Other	17	18	35
Portuguese	13	15	28
Chinese	9	4	13
Total	114	87	201

[a]The following languages had five or fewer speakers: Arabic, Armenian, Bambara, French, Haitian Creole, Italian, Japanese, Mauritian Creole, Mongolian, Spanish, Tajik, Thai, Ukrainian, and Vietnamese

had been in use for six semesters so the maximum number of semesters a rater could have rated was six. The expert panel consisted of eight raters with an average of 4.75 semesters of rating experience and a range of 3 to 6 semesters. For each prompt, the raters used the speaking score rubric to predict the level of language an examinee would need to succeed with the prompt identified what objectives were being measured, and provided feedback on whether they felt the item would function as intended.

The results of the ratings assigned by the eight raters were used to obtain the expert-predicted item difficulty (EID). The experts were presented 15 prompts, and based on their feedback, 10 prompts, two for each of the targeted levels, were selected for inclusion on the test. The items were designed with varying amounts of preparation time and response time to meet the functions of the prompt.

The EIDs of the 10 selected items rose monotonically in that every Level 2 prompt was easier than every Level 3 prompt and every Level 3 prompt was easier than the Level 4 prompts, etc. and this was considered to be evidence that the prompts did reflect the scale descriptors. The EIDs were also used to examine the extent to which the estimated difficulty of the speech prompts ordered as expected with the full-range and at-level rubric. The speaking test was designed with the same framework as an interview-based test that progresses from easier to harder and then back down so that examinees would experience a full range of prompt difficulties.

Rating and Scoring Procedures

The assessment was administered to students as part of their final exams. The scoring rubric was based on an 8-level scale that roughly corresponded to the ACTFL OPI scale. The rubric addressed three axes: (a) text type (e.g., word and phrase length, sentence length, paragraph length, etc.), (b) content, and (c) accuracy. Each axis ranged from *no ability* to *high ability* (i.e., the functional equivalent of a well-educated highly articulate native speaker). The scale was intended to be noncompensatory so that a response that was native-like in one area (e.g., pronunciation) could not compensate for a weak performance in another area (e.g., a text type that was only word length). The full-scale rubric required raters to keep the full range in mind as they judged performances. The at-level scale allowed raters to focus on a restricted-range of five levels: far below level, below level, at-level, above level, and far above level (see Table 2). To ensure the results had the characteristics of interval data and fully justified the use of parametric statistics, both ratings (holistic and analytic) were converted from raw scores (typically used in classical test theory) to fair averages (based on Rasch modeling).

Table 2 Full-range speaking rubric to at-level scale conversion matrix

	At-level	Intended item difficulty level				
	Rating	2	3	4	5	6
Below by 2 or more levels	1					0
					0	1
				0	1	2
			0	1	2	3
		0	1	2	3	4
Below by 1 level	2	1	2	3	4	5
At-level	3	2	3	4	5	6
Above by 1 level	4	3	4	5	6	7
Above by 2 or more levels	5	4	5	6	7	
		5	6	7		
		6	7			
		7				

Rating Methods

The tests were rated by judges with ESL training that were working as teachers. All of the raters had been trained at various times to use the rubric for the regularly scheduled computer-administrated speaking tests. The existing rater training material was designed to train raters how to use the 8-level scale for test-level scoring. The raters had received over 3 h of training and completed a minimum of 12 calibration practice ratings to ensure sufficient knowledge of the rubric. The raters had a packet that contained a copy of the rubric and a printed copy of the exam prompts, the objective of the prompt, and the intended difficulty level of the prompt. Details on how the test-level rubric was adapted will be discussed below.

Item-level rating designs. To get the item-level statistics, each test had to be rated at the item level. There were two possible incomplete connected design possibilities that could have provided the requisite data. The first was an *incomplete, connected* design in which all the items on a single test were rated by raters, who were linked to other raters. While this design is more cost-effective than a fully-crossed design, examinee ability estimates can be biased if there is an "unlucky combination of extreme raters and examinees" (Hombo et al. 2001, p. 20). The second design possibility was an *incomplete, connected spiral* design. This design was differentiated from the prior by assigning individual items to raters and linking raters to other raters through shared item ratings (Eckes 2011). This design shared the cost-effectiveness of the incomplete, connected designs, but has some distinct advantages. First, when raters listen to the same item from different examinees, they can have a deeper understanding of the response characteristics needed to assign a rating. Second, the spiral design can minimize errors associated with the *halo effect*. Halo effect occurs when performance on one item biases the

Table 3 Incomplete spiral connected design for analytically rated speaking test by prompt

Students	Prompt	Raters			
		1	2	3	4
1–4	1, 2, 3, 4	X	X	X	X
5	1	X	X		
5	2		X		
5	3		X	X	
5	4		X		X
6	1	X		X	
6	2		X	X	
6	3			X	
6	4			X	X
7	1	X			X
7	2		X		X
7	3			X	X
7	4				X

rating given on subsequent prompts (Myford and Wolfe 2003). For example, if a rater listens to a prompt and determines the examinee to speak at a Level 4 based on the rubric, then the rater might rate all subsequent prompts at 4 even when the performance might be higher or lower. Finally, spiral rating designs have been found to be robust in providing stable examinee ability estimates in response to rater tendencies (Hombo et al. 2001).

For this design, each rater was assigned to rate a single prompt (e.g., rater 1 scored all of prompt 1, rater 2 scored all of prompt 2, etc.). To avoid having disconnected subsets, a subset of the same students was rated on each item by all the raters. To further ensure raters were familiar with the items, raters rated some additional tests in their entirety. Table 3 presents an example of an incomplete, spiral design representing seven examinees, four raters, and four prompts. For the actual study, the design included 201 students, 10 raters, and 10 prompts.

Full-Range scale. Since all existing training materials for the rubric were designed in rating tests as a whole, the raters had to be given special instructions. They knew the intended level of the prompt they were scoring, and were told to reference that as they applied the rubric. For example, when rating a prompt that was designed to elicit Level 2 speech samples (ask simple questions at the sentence level), a rater was able to use the entire range of categories in the rubric (0–7). Since a rating of 2 would be passing, the only way the higher categories would be used is if the examinee spontaneously used characteristics of those higher categories through the use of more extended discourse, more academic vocabulary, native-like pronunciation, etc.

At-Level scale. To compensate for the possibility that a restricted-range bias or misuse of the rating rubric impacted the scores, the ratings were recoded to a five-point scale referred to as the at-level scale. Since the raters knew the intended level of the prompt they were evaluating, the rating likely reflected whether the

student response was below the targeted prompt level, at-level or above level. Table 2 shows how each intended level's 8-point rubric was converted to the 5-point at-level scale.

For example, with a Level 2 prompt, a rating of 0 which indicated little or no language on the holistic scale would be transformed to a 1, a rating of 1 which indicated that the language elicited for the Level 2 prompt was still below level was transformed to a 2, a rating of 2 which indicated that the language elicited was at-level and was transformed to a 3, a rating of 3 which indicated that the language elicited fulfilled all the required elements of level 2 language and had characteristics of Level 3 language was transformed to a 4, and a rating of 4, 5, 6 or 7 which indicated that the language elicited had characteristics of those levels was transformed to a 5. Similar conversions were calculated for each of the prompts. Thus, if the response was deemed at level for that prompt the associated score would be a 3.

Data Analysis

To answer the questions on how reliable the two item-level scales functioned, the facets software was used to conduct Many Facets Rasch modeling (MFRM). In MFRM, the facets were modeled in such a way that person ability, item difficulty, and rater severity were measured conjointly with the rating scale.

Measurement invariance. Besides being interval data, another advantage of using Rasch scaling is that the parameter estimates for both persons and items have the quality of *measurement invariance* (Engelhard 2008). That is, when measuring a unitary construct, person ability estimates are the same regardless of the items that are presented to the examinees, and item ability estimates are the same regardless of the examinees who respond to them. Since the application of the findings of this study were directed for test developers in equating test forms, measurement invariance of the items was highly relevant. Beyond the advantage of measurement invariance, the Rasch analysis provided information on how well the scale functioned and the reliability of the test scores and test items.

Diagnoses of rating scales. To evaluate how well a scale functions with Rasch measurement, there are a number of diagnostics available including (a) category frequencies, (b) average logit measures, (c) threshold estimates, (d) category probability curves, and (e) fit statistics (Bond and Fox 2007). For category frequencies, the ideal is that there should be a minimum of 10 responses in each category that are normally distributed. For average logit measures, the average person ability estimate of each rating category should increase monotonically (Eckes 2011). The threshold estimates are the logits along the person ability axis at which the probability changes from a person being in one category to another. Those estimates should increase monotonically as well. In order to show distinction between the categories, they should be at least 1.4 logits apart and to avoid large gaps in the variable and the estimate should be closer than five logits (Linacre 1999). When looking at a graph of the category probability curves, each curve

should have its own peak, and the distance between thresholds should be approximately equal. If one category curve falls underneath another category curve or curves, then the categories could be disordered and in need of collapsing. Finally, fit statistics provide one more way to examine a rating scale. If the outfit mean squares of any of the categories are greater than 2.0, then there might be noise that has been introduced into rating scale model (Linacre 1999). Using these diagnostics through a FACETs analysis, a measurement scale can be analyzed.

Reliability analysis. Finally, Rasch scaling provides more tools in determining the reliability of test scores, especially when there are multiple facets. Reliability is defined as the ratio of the true variance to the observed variance (Crocker and Algina 1986). Unlike classical test theory which can only report reliability on the items of a test (e.g., Cronbach's alpha or Kuder–Richardson 20) or the agreement or consistency of raters (e.g., Cohen's kappa or Pearson's Correlation coefficient), Rasch reliability reports the relative reproducibility of results by including the error variance of the model in its calculation. Furthermore Rasch reliability provides estimates for every facet (person, rater, item) that is being measured. When the reliability is close to 1.0, it indicates that the observed variance of whatever is being measured (person, rater, item) is close or nearly equivalent to the true (and immeasurable) true variance. Therefore, when person reliability is close to 1, the differences in examinee scores are due to differences in examinee ability. If there are multiple facets such as raters, it might be desirable for a construct irrelevant facet to have a reliability estimate close to 0. If raters were the facet, a 0 would indicate the raters were indistinguishable from each other and therefore interchangeable. Any examinee would likely obtain the same rating regardless of which rater were assigned to them. Conversely, if the rater facet had a reliability estimate close to 1.0, then the raters are reliably different and the rating obtained by a given examinee is highly dependent on the rater. When the rater facet is not close to 0, it is necessary that an adjustment be made to the examinee score to compensate for the rater bias.

To obtain item-level ratings, the item speaking level was calculated using a FACETS analysis. The three facets for this analysis included examinees, raters, and prompts. Since the raters used (a) a full-range 8-point scale that could be converted to (b) a restricted-range 5-point at-level scale at the prompt level, the Andrich Rating Scale model was the most appropriate to use (Linacre and Wright 2009). One of the requirements for the use of Rasch MFRM is that the data be unidimensional and while it may appear that that a noncompensatory scale with three axes by definition does not have that characteristic, we argue that true unidimensionality rarely occurs in educational measurement and that essential unidimensionality can exist through shared understanding of construct definitions (see Clifford and Cox 2013) and the fit of the data (McNamara and Knoch 2012).

The rubrics used for the rating scale were (a) the same as the holistic rated speaking level scale but applied to individual prompts (or items) on the exam and the derived at-level scale (see Table 2). To see which rubric (full-range or at-level) most closely aligned with the expert-predicted item difficulty (EID) of each prompt, the item difficulty parameters from the two rubrics were correlated with the EIDs.

Results

To answer the research questions of this study on scale reliability required us to obtain scores at the prompt level. An 8-level full-range rubric that is typically used to rate at the test-level raters was applied to each prompt of the assessment. The ratings were then converted to an at-level scale that evaluated if the examinee performed below, at or above the intended difficulty level. A facet analysis was conducted with the ratings from the full-range and at-level scale. To do a comparison between the EID of the prompts and the actual difficulty based on ratings at the prompt level, a correlation was run.

Full-Range Scale Rasch Analysis. The items were rated with the full-range scale using an incomplete, spiral connected design and scale and reliability analyses were conducted.

Scale analysis. The eight categories of the full-range rubric functioned within acceptable parameters for the study (Table 4). With the exception of the category 0 (n = 3), the relative frequency of each category had a minimum of 10 in each category. The average measure increased monotonically from 1 to 7 without exception, as did the threshold estimates. The threshold estimates had the minimum recommendation of 1.4 logits between each category indicating that each category showed distinction, and none of the thresholds were over 5 logits apart, and the spacing of the categories was more evenly spaced than the human-rated holistic speaking level. An examination of the category probability distributions was indicative that each category functioned well (see Fig. 1).

For the fit statistics, the outfit mean squares of the categories did not exceed 2.0 with the exception of the 0 category which only had 3 responses. The only category that did not fit the guidelines of a good scale was 0. Since the 0 category is typically reserved for little or no production and since the students had one semester of instruction, this category could be combined with category 1 if it were only used as an end of instruction scale. However, since the scale is used for placement testing as well, all eight categories were retained. The full-range rubric functioned within acceptable parameters to be used in the analysis.

Table 4 Full-range rubric scale category statistics

Category	Absolute frequency	Relative frequency (%)	Average measure	Outfit	Threshold	SE
0	3	0	−3.04	2.0		
1	54	2	−3.72	1.1	−7.13	0.59
2	345	12	−2.39	1.0	−5.06	0.77
3	933	33	−0.23	1.0	−2.27	0.27
4	916	32	1.62	0.9	0.77	0.16
5	449	16	2.91	1.0	3.00	0.17
6	142	5	3.88	1.1	4.55	0.23
7	24	1	4.93	1.0	6.14	0.49

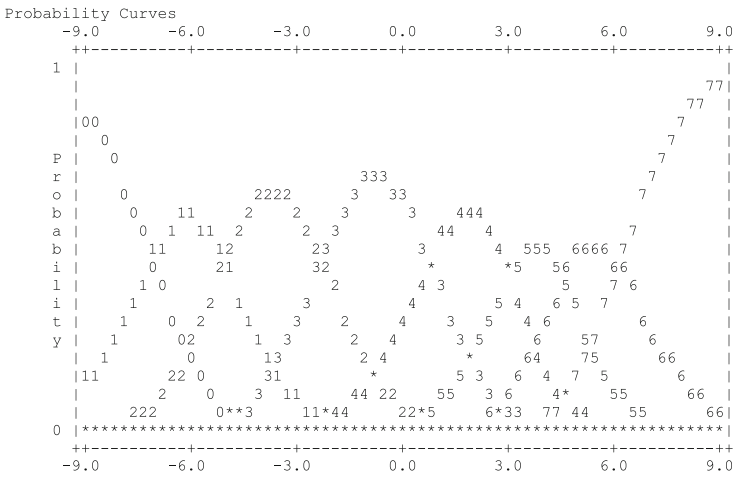


Fig. 1 Full-range rubric rating category distribution

Reliability analysis. The reliability statistics with the full-range rubric found that all three facets were reliably separated. Figure 2 is a vertical scale map that shows the logit in the first column, the examinee ability level in the second column, the rater severity in the third column, the empirical item difficulty in the fourth column and the scale equivalency in the fifth column. The 0 in the middle of the vertical scale is tied to the means of the rater and item ability estimates or logits. An examinee with an ability logit of 0 (the second column) would have a 50 % chance of being rated in category 3 (the fifth column), by raters R5 or R9 (the third column) on item L4-2 or L6-2.

Figure 2 showed that the examinee ability ranged from category 2 to 6. The examinees separation reliability was 0.94, thus we can be confident of the different ability levels of the examinees. For rater reliability, there was a separation reliability coefficient of 0.96. In Fig. 2, we can see that the raters R7 and R10 were the most generous and rater 4 was the most severe. Even though the raters rated the items differently than one another, the fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0. Thus, the rater severity error could be mathematically modeled out of the examinees' scores through the use of the fair average.

The item facet had a reliability of 0.89 indicating that items could not be used interchangeably without compensating for their difficulty level. In Fig. 2, the third column represents the intended level and item number. The easiest item was L6-1 (i.e. EID Level 6, Item 1) and the most difficult item was L5-1. While it was expected that the prompts would have varying item difficulties and that some kind of item equating would need occur to create equivalent test forms, it was unexpected that the item difficulty means did not order in their intended levels. This could be due to a restricted-range error in which raters were unwilling to use the extremes of the rubric. It was also notable that the prompts clustered around

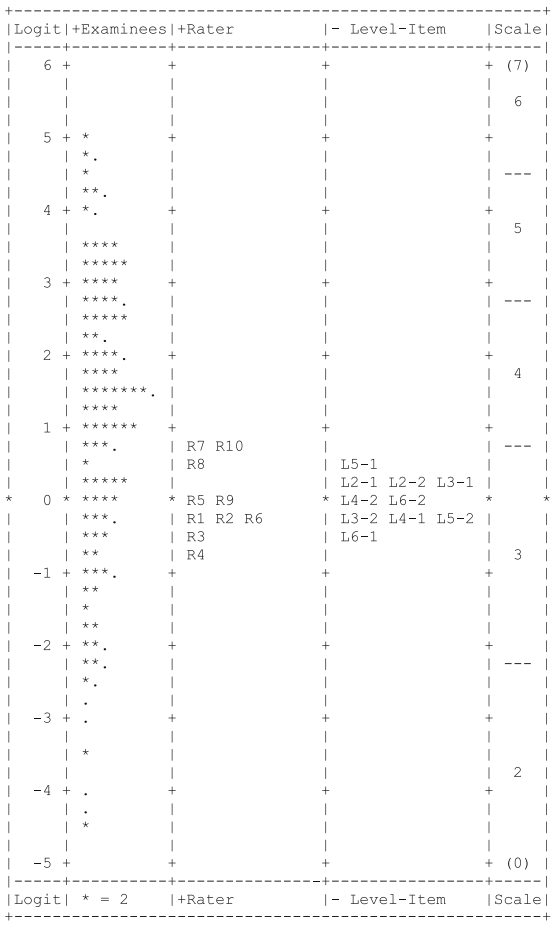


Fig. 2 Analytic full-range speaking level vertical scale

category 3 (SD = 0.27) and had a narrower range than the raters (SD = 0.48). The prompt fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0.

At-Level Scale Rasch Analysis. The item ratings from the full-range scale were converted to the at-level scale and analyzed with FACETS.

Scale analysis. The at-level scale functioned within the parameters needed for a reliable scale (see Table 5). The relative frequency of each category had a minimum of 10 in each category. The average measure increased monotonically without exception, as did the threshold estimates. The threshold estimates had the minimum recommendation of 1.4 logits between each category indicating that each category showed distinction, and none of the thresholds were over 5 logits apart. Furthermore, the spacing between the thresholds was more evenly spaced than the full-range scale (see Fig. 3). The outfit mean squares of the other categories did not exceed 2.0.

Table 5 At-level scale rating scale category statistics

Category	Absolute frequency	Relative frequency (%)	Average measure	Outfit	Threshold	SE
1	770	27	−6.14	1.0		
2	553	19	−2.52	0.6	−3.86	0.08
3	589	21	−0.06	1.0	−1.39	0.07
4	552	19	2.38	1.2	1.19	0.07
5	402	14	5.22	1.1	4.07	0.09

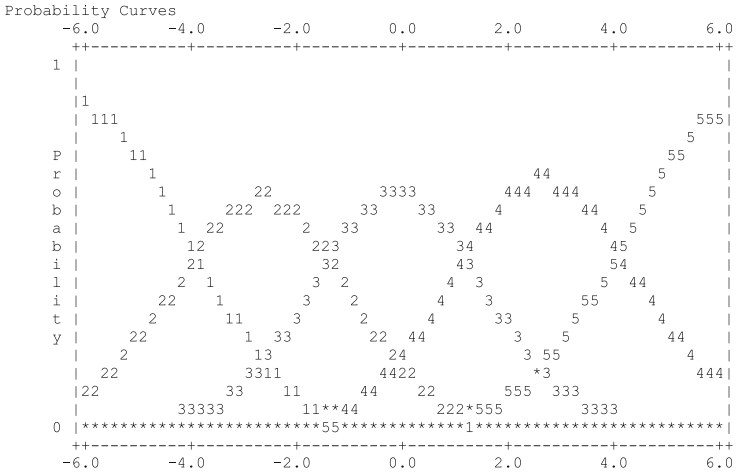


Fig. 3 At-level scale rating category distribution

Reliability analysis. The reliability statistics on the at-level item scoring found that all three facets (examinees, raters, and items) were reliably separated. In Fig. 4, we can see that the examinees have a range from categories 1–5. Note that the significance of these categories did ***not*** signify the levels of the speaking rubric facets of examinee, but rather whether how well they performed the task at its intended level. This analysis found that the separation reliability between the examinees was 0.93 and that the examinees could be separated reliably into different groups.

The raters had a reliability of 0.96 the same as the full-range analysis with the standard deviation being slightly larger than the at-level scale analysis (full-range rubric SD = 0.55 compared to at-level scale SD = 0.49). The raters still exhibited different levels of severity with R7 being the most generous and raters R3 and R4 being the most severe. Comparing the raters in Figs. 2 and 4 we see that the ordering of the severity and generosity of the raters is very similar with a high correlation ($r = 0.78$, $p < 0.05$) between the rating scales. The fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0.

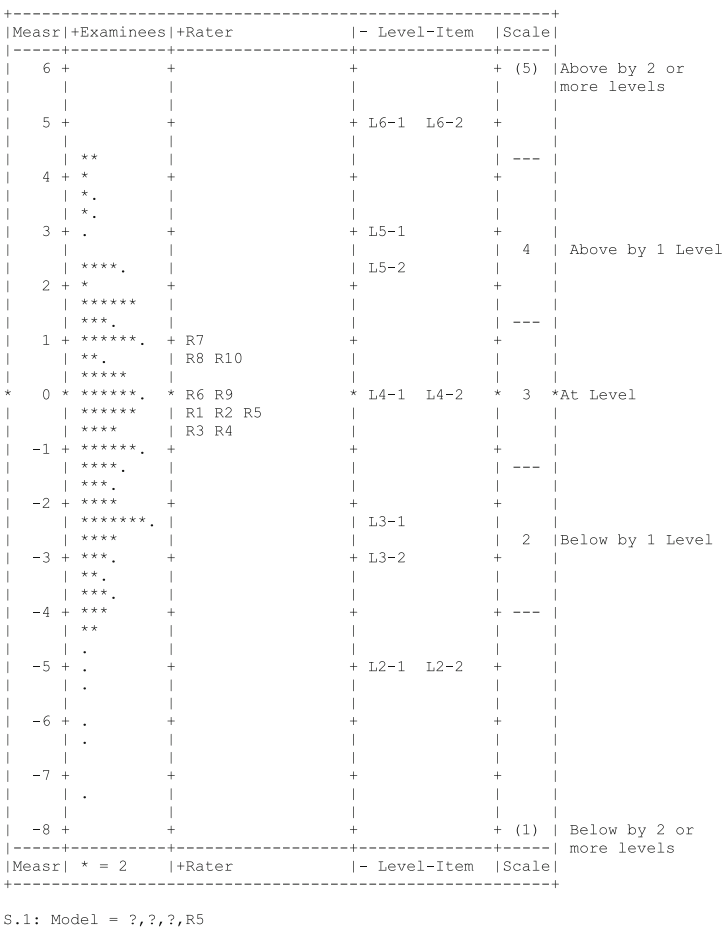


Fig. 4 At-level vertical scale of examinees, raters, and items

Most noteworthy though was the fact that the at-level scale item facet jumped to a reliability of 1.00 indicating that it would be virtually impossible to have the same score with the different items. Furthermore, the differences in difficulty aligned closely with the EID (see Fig. 5). The prompts that were intended to elicit Level 6 language (L6-1 and L6-2) were the most difficult while the prompts intended to elicit Level 2 language were the easiest (L2-1 and L2-2). Such high item separation reliability is in part due to the at-level scale being honed into the level that the prompt was eliciting. In other words, the only way for a prompts targeted at 2 subsequent levels to be conflated would be for the lower level prompt to have a preponderance of 4 and 5 ratings and the higher level prompt have a preponderance of 1 and 2 ratings. Table 6 illustrates the manner in which the item difficulty parameters increase monotonically as the intended difficulty increased.

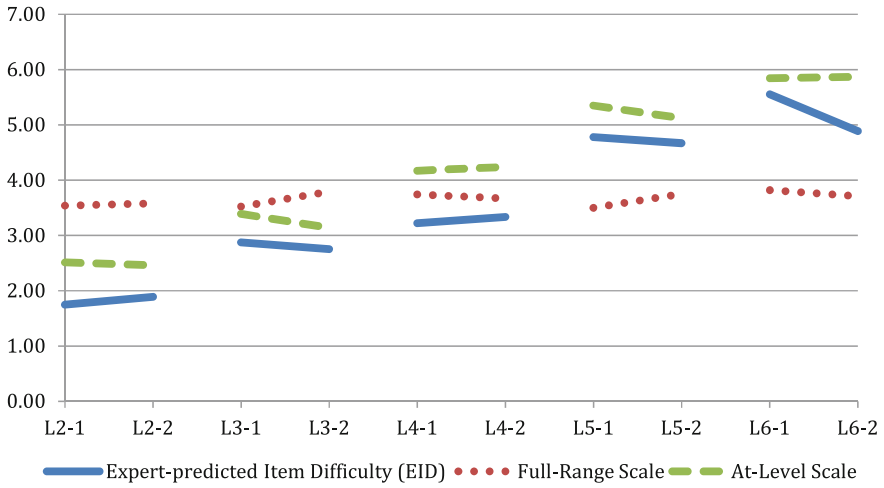


Fig. 5 Means of item difficulty measures

Table 6 At-level scale item statistics in order of measure

Item	Fair average	Logit	[a]Conversion to full-range rubric
L2-1	4.49	−4.86	2.51
L2-2	4.54	−5.04	2.46
L3-1	3.61	−2.22	3.39
L3-2	3.86	−2.93	3.14
L4-1	2.83	−0.15	4.17
L4-2	2.76	0.02	4.24
L5-1	1.65	2.92	5.35
L5-2	1.88	2.27	5.12
L6-1	1.16	4.88	5.84
L6-2	1.13	5.10	5.87
Mean	2.79	0.00	4.21
S.D.	1.30	3.74	1.30

[a]The conversion to the full-range rubric consisted of subtracting the Fair Average from seven, which was the highest category level of the full-range rubric

Comparing Prompt Difficulty

To determine the extent to which the expert-predicted item difficulty aligned with the full-range and at-level rubrics, the item difficulty statistics for the ratings of both rubrics were calculated. The full-range scale spanned from 0 to 7, and the averages of all three measures were in the middle of the scale with the ratings range from

Table 7 Comparison of speaking level item statistics

Item	[a]Expert predicted item difficulty	Full-range rubric fair average by prompt	At-level rubric converted fair average
L2-1	1.86	3.54	2.51
L2-2	2.00	3.58	2.46
L3-1	2.71	3.52	3.39
L3-2	2.71	3.79	3.14
L4-1	3.13	3.74	4.17
L4-2	3.38	3.67	4.24
L5-1	4.63	3.50	5.35
L5-2	4.75	3.75	5.12
L6-1	5.00	3.82	5.84
L6-2	5.50	3.71	5.87
Mean	3.57	3.66	4.21
S.D.	0.71	0.11	1.30

[a]Based on average from eight different expert raters

Table 8 Post-study correlations between item difficulty measures

	At-level rubric item difficulty	Full-range rubric item difficulty
Expert-predicted item difficulty	0.98	−0.43
At-level rubric item difficulty		−0.42

Note Correlations greater ±0.77 are significant at $p < 0.01$ (2-tailed)

3.50 to 3.82 (see Table 7), while the at-level scale had item ratings range from 2.51 to 5.87 (see Fig. 5).

A Pearson Product moment correlation was run between all of the item measures (see Table 8): the expert-predicted item difficulty, the full-range scale and the at-level scale. The highest correlation was between the at-level scale and the expert-predicted difficulty ($r = 0.98$, $p < 0.001$), but the full-range scale had a slight inverse relationship ($r = -0.43$) with the EID.

Discussion and Conclusions

Both the full-range and at-level rubrics functioned well as rating scales. With the exception of the 0 category with the full-range scale, each of the categories of both scales were used with enough frequency that there was no need to combine adjacent categories. Both of the scales resulted in high separation statistics between examinees, raters and items. The separation between examinees is desirable, but the

separation between raters could be cause for concern. The decision on how to treat rater disagreements depends if they are to be treated as "rating machines" that are interchangeable with one another or if they are to be viewed as "independent experts" that are self-consistent but whose ratings must be mathematically modeled and transformed to provide examinees with a fair score (Linacre 1999). It is often assumed that enough training can force raters to act as machines, however, without careful follow-up, that may not be the case (McNamara 1996) and acknowledging the disagreements of independent experts through mathematical modeling can more fully reflect real-world rating circumstances (Eckes 2011).

While both the full-range and at-level scales resulted in item difficulty statistics that were statistically separated, the ordering of the full-range rubric difficulties did not align with the experts' predictions. There are a number of possible explanations as to why the first could be that the descriptors in the scale upon which the items were written were flawed. While possible, the scale was based on the well-established ACTFL scoring rubric that has been in use for over 30 years, and after the 1999 revision was validated in inter-rater consistency in over 19 languages (Surface and Dierdorff 2003). Second, there is the possibility that the items did not adequately reflect the scales' descriptors. The raters that evaluated the prompts to determine their intended difficulty levels had a minimum of 3 semesters of rating experience with the average number of semesters being 4.75 semesters. These raters felt that the items did align with the rubric they used for rating.

Another possibility is likely the existence of a pervasive restricted-range error in using a full-range scale to rate a single item. When an item is targeted at Level 6, and the rater knows it is targeted at Level 6, that rater might be hesitant to give scores on the lower end of the scale (0, 1, and 2) even if the respondent language is characteristic of those levels. Similarly an item targeted at Level 2 might result in ratings that are not in the higher part of the range (5, 6, and 7) because the prompt did not elicit language in that upper range. This range restriction in scoring could have resulted in fair item averages that clustered close to the mean of all the items. One piece of evidence of this possibility is the fact that full-range ratings had the smaller fair average standard deviation (see Table 7) of the two scoring methods.

One additional explanation is that that raters did not adhere to the rubric when scoring responses provided by individual at the individual prompt level. Rather raters may have scored the response to each prompt based on how well they answered and not whether the response provided evidence that the examinee was able to speak well at that level. For example, the content of an examinee's response may have been very interesting, yet the language produced did not have the requisite abstract vocabulary and command of more complex grammar needed for a higher level rating. In this instance, the rater may have awarded a higher score than justified by the defined categories of the rubric. This may be the case as it is unlikely that a prompt intended to elicit a response demonstrating the examinees ability to speak at a basic level would consistently provide evidence that the examinee was able to speak at a higher level.

The use of the at-level scale, however, allowed for the EIDs to align with empirical item difficulties. Another benefit of this scale was that it gave information on prompts that were intended to elicit language at the same level. With the items in this study, for example, the prompts at Levels 2, 4 and 6 could be used interchangeably with the other prompts at those intended levels because their item difficulty parameters had comparable values. The prompts at Levels 3 and 5 however were not comparable. Item L5-1 was more difficult than Item L5-2 and similarly item L3-1 was more difficult than Item L3-2. If there were more prompts, then test developers could chose those that would create equivalent test. The prompt fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0.

Discussion and Review of Findings

The research question this study addressed explored how the application of a scoring rubric (full-range and at-level) affected the reliability of the results as well as how the two rubric applications compared with the expert-predicted item difficulties. The implications of these findings impacts how to best create equivalent test forms for speaking exams. In creating a speaking proficiency test that is tied to a set of standards, an item writer would try to elicit a specific proficiency level of speech in the construction of the prompt. Consider the following prompts: (1) Describe your house and the neighborhood you live in, and (2) What is the impact of government subsidized housing on the quality of life in a neighborhood?

In the first prompt, the intent is to elicit speech at the ACTFL Intermediate level, whereas in the second prompt, the intent is to elicit speech at the Superior level. If those intended prompt difficulties do not align with the empirical human rating, there are important implications for item writers attempting to create parallel test forms. The determination of item equivalence from one test form to the next needs to be justified by demonstrating that item writers can reliably write prompts to an intended difficulty level.

Training raters on a full-range scale would be ideal for many reasons. First, they would have an understanding of the entire range of a set of standards and see how any performance relates to those standards. Feedback on examinee performance could be easily provided to examinees, teachers and other stakeholders and it could occur independent of the task presented to the examinee. Unfortunately the ratings of examinee responses at the prompt level using a full-range scale did not align with the EID levels. First, the item difficulty statistics had very little variance (SD = 0.27). In fact, Fig. 2 illustrated that the difference in the raters was greater than that of the items (SD = 0.48). Furthermore, the correlation between the EID levels and the full-range item fair averages were not statistically significant and inversely correlated ($r = -0.43$). The incongruence of trained raters (a) being able to predict differences in prompt difficulty yet (b) being unable to find performance differences from the prompts leads one to question the full-range rubric rating

approach. Using the holistic 8-level scale on each item seemed to have introduced a restricted-range error. This could be an artifact of telling the raters the intended level of the prompt they were rating, but it could also be failure to use the rubric properly. Raters more likely scored the responses for each item based on how well they provided evidence the examinee responded to the prompt.

The full-range rating scale spanned a range of language possibilities from simple sentences on familiar topics (Level 2) to extensive, complex speech on abstract academic topics (Level 6), but each of the individual prompts was aimed at only one of those levels (i.e., each prompt was intended to elicit evidence of speaking ability at a specific level and not higher). This misalignment created challenges and perhaps even cognitive dissonance for the raters. For instance, it would be difficult for a prompt targeted at a Level 2 task to elicit a speech sample much higher than a Level 3 or at most a Level 4, even if the examinee did respond with more extensive speech. The rater might be reticent in awarding a rating that was more than 2 levels higher than the prompt's intended difficulty level. Conversely, when a rater was scoring a failed attempt at a prompt targeted at Level 6 task, it might be difficult to know why an examinee was failing to perform at that level and there might be little evidence about what level the examinee could accomplish. The failure to offer an academic opinion on complex topics could mean the examinee was a beginning speaker with almost no speaking ability or it could be an intermediate speaker suffering linguistic breakdown because of the increased cognitive load. Raters might not know how low to rate such breakdown and may be reticent to assign a rating more than 2 or 3 levels below the prompt's intended difficulty level. Thus the ratings for all of the prompts judged with the full-range rubric clustered around the mean (mean = 3.66, SD = 0.11).

Using an at-level scale for each item (through the conversion of the holistic rubric ratings) functioned much better from a measurement perspective. First, there was a wider dispersion of the prompt difficulty means (mean = 2.79, SD = 1.30). Second, the Rasch analysis showed the categories had the most uniform distribution so the categorical differences in ratings examinees received were the most equidistant (see Fig. 3). Finally, there was a much stronger relationship ($r = 0.98$, $p < 0.01$) with the expert-predicted difficulties (see Table 8) than there had been with the holistic scale ratings. Therefore, the low relationship established through using the full-range scale at the item level could be more indicative of scale misuse than the inability of the raters to differentiate performance when judging the different prompts. The result seems to indicate that either responses obtained from lower level prompts did provide some evidence of the examinee's ability to speak at a higher level or that raters tended to rate the respondents overall quality of the response on an 8-level scale. Either way, an analysis of the at-level scale data verifies that the intended prompt difficulty did affect the overall assessment of speaking ability. From the result of this analysis it was also noted that those prompts intended to elicit evidence of speaking ability at Levels 2 (L2-1, L2-2), 4 (L4-1, L4-2), and 6 (L6-1, L6-2) were of equal difficulty within level, but the prompts at Levels 3 (L3-1, L3-2) and 5 (L5-1, L5-2) were not of equal difficulty. Analyzing prompts in this way can provide evidence of test equivalence when

attempting to create parallel forms of an assessment. It also provides an item-level statistic of difficulty that could be used when creating item banks. Finally, since the at-level scale is a subset of the full-range, it can maintain many of the advantages of the full-range scale through simple mathematical conversion.

Limitations and Future Research

In this study, the raters who judged the item responses had been trained to rate overall performances with a holistic scale. They were not given any instruction or exemplars on how to apply the scale at the item level, and the task may have been untenable, as the rubric was not designed for use at the micro level of item. This design weakness might have been overcome if there had been more rater training that focused on the item level. Through the training, the challenge of implementing a holistic scale at the item level could have emerged and a change to the design could have been implemented at that time. Fortunately, the existing holistic scale could be converted after the fact so a more accurate analysis could still be made. Were the research to be done again, it would be better to (a) initially design the scale at the item level and (b) trial the scale to ensure it functions as intended. This would have avoided the step of needing to conduct a post hoc analysis.

Treating a rubric with three distinct axes (text type, content, and accuracy) as a unidimensional construct could have affected the rating as well. Raters making expert judgments of performance have a cognitive load placed upon them that could be simplified by letting them focus only on one aspect at a time. Then, if a multidimensional IRT model had been applied, the findings might have been different as well.

Conclusions

Using full-range rubrics to rate individual items that are targeted at specific levels is problematic and should be done only with caution and verification that the ratings are free from rater error (central tendency, range restriction or logical errors). In this study, a 5-point at-level scale derived from a full-range application of the rubric but targeted at the intended level of the prompt yielded much better results. Using this scale, prompts that were targeted to elicit speech at the same level were more likely to represent their intended empirical difficulty levels. There was a clear separation in the scoring based on the intended prompt difficulty levels which would allow for these data to be used when creating equivalent forms of a test or selecting items for adaptive testing.

Appendix A

Speaking Rubric

Level	Text Type	Accuracy	Content
7—leaving Academic C	Exemplified speaking on a paragraph level rather than isolated phrases or strings of sentences. Highly organized argument (transitions, conclusion, etc.). Speaker explains the outline of topic and follows it through.	• Grammar errors are extremely rare, if they occur at all; wide range of structures in all time frames; • Able to compensate for deficiencies by use of communicative strategies—paraphrasing, circumlocution, illustration—such that deficiencies are unnoticeable; • Pausing and redundancy resemble native speakers; • Intonation resembles native-speaker patterns; pronunciation rarely if ever causes comprehension problems; • Readily understood by native speakers unaccustomed to non-native speakers.	• Discuss some topics abstractly (areas of interest or specific field of study); • Better with a variety of concrete topics; • Appropriate use of formal and informal language; • Appropriate use of a variety in academic and non-academic vocabulary.
6—starting Academic C	Fairly organized paragraph-like speech with appropriate discourse markers (transitions, conclusion, etc.) will not be as organized as level 7, but meaning is clear.	• Grammar errors are infrequent and do not affect comprehension; no apparent sign of grammatical avoidance; • Able to speak in all major time frames, but lacks complete control of aspect; • Pausing resembles native patterns, rather than awkward hesitations;	• Uses appropriate register according to prompt (formal or informal); • Can speak comfortably with concrete topics, and discuss a few topics abstractly; • Academic vocabulary often used appropriately in speech.

(continued)

(continued)

Level	Text Type	Accuracy	Content
		• Often able to successfully use compensation strategies to convey meaning.	
5—starting Academic B	Simple paragraph length discourse.	• Uses a variety of time frames and structures; however, speaker may avoid more complex structures; • Exhibits break-down with more advanced tasks—i.e. failure to use circumlocution, significant hesitation, etc; • Error patterns may be evident, but errors do not distort meaning; • Pronunciation problems occur, but meaning is still conveyed; • Understood by native speakers unaccustomed to dealing with non-natives, but 1st language is evident.	• Able to comfortably handle all uncomplicated tasks relating to routine or daily events and personal interests and experiences; • Some hesitation may occur when dealing with more complicated tasks; • Uses a moderate amount of academic vocabulary.
4—starting Academic A	Uses moderate-length sentences with simple transitions to connect ideas. Sentences may be strung together, but may not work together as cohesive paragraphs.	• Strong command of basic structures; error patterns with complex grammar; • Pronunciation has significant errors that hinder comprehension of details, but not necessarily main idea; • Frequent pauses, reformulations and self-corrections; • Successful use of compensation strategies is rare; • Generally understood by sympathetic speakers accustomed to speaking with non-natives.	• Able to handle a variety of uncomplicated tasks with concrete meaning; • Expresses meaning by creating and/or combining concrete and predictable elements of the language; • Uses sparse academic vocabulary appropriately.

(continued)

(continued)

Level	Text Type	Accuracy	Content
3—starting Foundations C	Able to express personal meaning by using simple, but complete, sentences they know or hear from native speakers.	• Errors are not uncommon and often obscure meaning; • Limited range of sentence structure; • Intonation, stress and word pronunciation are problematic and may obscure meaning; • Characterized by pauses, ineffective reformulations; and self-corrections; • Generally be understood by speakers used to dealing with non-natives, but requires more effort.	• Able to successfully handle a limited number of uncomplicated tasks; • Concrete exchanges and predictable topics necessary for survival; • Highly varied non-academic vocabulary.
2—starting Foundations B	Short and sometimes incomplete sentences.	• Attempt to create simple sentences, but errors predominate and distort meaning; • Avoids using complex/difficult words, phrases or sentences; • Speaker's 1st language strongly influences pronunciation, vocabulary and syntax; • Generally understood by sympathetic speakers used to non-natives with repetition and rephrasing.	• Restricted to a few of the predictable topics necessary for survival (basic personal information, basic objects, preferences, and immediate needs); • Relies heavily on learned phrases or recombination of phrases and what they hear from interlocutor; • Limited non-academic vocabulary.
1—starting Foundations A	Isolated words and memorized phrases.	• Communicate minimally and with difficulty; • Frequent pausing, recycling their own or interlocutor's words; • Resort to repetition, words from their	• Rely almost solely on formulaic/memorized language; • Very limited context for vocabulary; • Two or three word answers in responding to questions.

(continued)

(continued)

Level	Text Type	Accuracy	Content
		native language, or silence if task is too difficult; • Understood with great difficulty even by those used to dealing with non-natives.	
0—starting foundations prep.	Isolated words.	• May be unintelligible because of pronunciation; • Cannot participate in true conversational exchange; • Length of speaking sample may be insufficient to assess accuracy.	• No real functional ability; • Given enough time and familiar cues, may be able to exchange greetings, give their identity and name a number of familiar objects from their immediate environment.

References

ACTFL Proficiency Guidelines 2012 (n.d.). In *American Council on the Teaching of Foreign Languages Proficiency website*. Retrieved from http://actflproficiencyguidelines2012.org/.

Bargainnier, S. (2004). Fundamentals of Rubrics. In D. Apple (Ed.), *Faculty Guidebook*. Lisle, IL: Pacific Crest Inc.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Buck, K., Byrnes, H., & Thompson, I. (1989). *The ACTFL oral proficiency interview tester training manual*. Yonkers, NY: ACTFL.

Clifford, R., & Cox, T. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals*, *46*(1), 45–61. Retrieved from Google Scholar.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.

Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement, 6*(3), 155–189.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation [Research Report (RR-01-05)]*. Retrieved from Google Scholar: Educational Testing Service.

Linacre, J. M. (1999). *A user's guide to FACETS*. Chicago: MESA press.

Linacre, J. M., & Wright, B. D. (2009). *A user's guide to WINSTEPS*. Chicago: MESA press.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

National Governors AssociationCenterfor Best Practices & Council of Chief State School Officers. (2010). *Common core state standards*. Washington, DC: Authors.

Tierney, R. & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, 9(2). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=2.

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., & North, B. (2009). Common European framework of reference for languages: Learning, teaching, assessment. ISBN:0521005310.

Determination of the Primary School Cooks' Knowledge, Attitude and Practice in Preparing Healthy School Meal Using Rasch Analysis

Zuraini Mat Issa and Wan Abdul Manan Wan Muda

Introduction

Previous studies have proven that consumption of nutritious meal provides a significant role in improving health, reduce morbidity, and give positive effect on school attendance and learning abilities (Abotsi 2013; Adelman et al. 2008; Hamid Jan et al. 2011; Monir et al. 2013; Ni Mhurchu et al. 2010, 2012; O'Neil et al. 2014). However, to ensure continuous healthy eating habit, the introduction to this practice should be started from childhood. Schools have been a perfect place to introduce nutritious meal and also to nurture healthy eating habit among the schoolchildren (Moore et al. 2013; Vecchiarelli et al. 2006; Wharton et al. 2008). It is also important to stress the need of having breakfast before school starts. The school community which includes the teachers, peers, and food operators including the cooks, should portray positive attitude and encouragement on good eating habit (Aznita Izma et al. 2009; Izumi et al. 2010). The school cooks should only cook and serve healthy meals at schools which later will be consumed by the school community including teachers, school staffs, and also the school children. However, there is a very limited study done in assessing the nutrition level of knowledge,

Z.M. Issa (✉)
Department of Foodservice Management, Faculty of Hotel and Tourism Management, Universiti Teknologi Mara, Kampus Puncak Alam, 42300 Bandar Puncak Alam, Selangor, Malaysia
e-mail: zurainim@salam.uitm.edu.my

W.A.M.W. Muda
School of Health Sciences, Universiti Sains Malaysia, Kubang Kerian Campus, 15150 Kubang Kerian, Kelantan, Malaysia
e-mail: wanmanan@usm.my

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_15

attitude and practice (KAP) of school cooks in preparing foods. Previous studies conducted by Arop et al. (2000, 2003), and Wan Nudri and Abdullah (1996) revealed that foods prepared and provided to the schoolchildren in primary schools in Malaysia are of low nutritional quality. Similar findings were reported by Nik Rosmawati et al. (2014), where most foods sold at schools in Kelantan are unhealthy. The foods were regarded as having too much carbohydrate, fat, protein, and added sugar, but very limited option for fruits, vegetables, and milk products. Therefore, the objective of this study was to examine the KAP level of the school cooks' in preparing healthy meals, and also to investigate the correlation between the KAP domains.

Methodology

HMP-KAP Instrument

A HMP-KAP instrument adapted from various sources where for the knowledge domain, the items were adapted from National Coordinating Committee of Food and Nutrition survey (NCCFN) (MOH 1997) and Chen et al. (2012), while the attitude items were adapted from Chen et al. (2012) and Aung et al. (2012). The items for practice domain were adapted from Chen et al. (2012) and Guidelines to Healthy School Canteen Management (MOE 2011). The knowledge items comprised of 15 multiple choice items and 10 were "true" or "false" or "not sure" items. During the analysis, the "not sure" responses were treated as missing data. The attitude domain comprised nine items with 4-point rating scales while the practice domain consists of 13 items with 3-point rating scales. The HMP-KAP instrument was validated prior to the study.

Data Collection

Before the instruments were distributed to all the respondents, an invitation letter was sent to all eligible school cooks via the respective District Office of Education in Kelantan, Malaysia. Out of 10 districts, nine of them responded to the letter. In total, 301 school cooks from 245 schools in Kelantan participated in the study. The data collection was conducted during weekends at the respective district office beginning March 2014 to mid of April 2014.

The instruments were then inspected for completion before analyzing the data using Winstep 3.80.1 for Rasch analysis and SPSS version 16 for other statistical

tests. The following criteria were used as a quality control to make sure the data provided were meaningful (Bond and Fox 2007):

- Positive Point-measure correlation (PTMEA Corr)
- The Infit and Outfit Mean Square should be between 0.5 and 1.5
- The Infit and Outfit ZSTD should be between −2.00 and 2.00
- Unidimensionality that should indicate either it is a strong (>40 %), moderate (>30 %) or minimal (>20 %) dimension (Conrad et al. n.d.).

The data were then subjected to further analysis. To further classify the school cooks according to categories proposed by the Technical Working Group on Research of Ministry of Health, Malaysia. According to the group, a good food handler should score at least 75 % for knowledge and attitude domains, and between 50 and 74 for a moderate score. On the contrary, food handlers are considered have a good practice if they score at least 50 % for the practice domain. If they score less than 50 % regardless of the domain, they are then considered of having either poor knowledge, attitude or practice. The logit measures were mapped to the proposed bench mark (Khatimin et al. 2013). The steps involved are as follows:

1. The items measure for each domain were sorted according to their difficulty level where the one with the lowest logit measure was considered as the easiest and thus it was expected that all respondents had the highest possibility of getting it right
2. If the cut-off point for 75 % bench mark is needed, and the number of items is 25 (example: knowledge domain), then the item that would correspond to the 75 % correct answers given would be the nineteenth item from the bottom (the easiest)
3. The logit measure for the nineteenth item would then be concluded as the cut-off point to get the 75 % and above
4. The same procedures apply to all the bench mark setting for all domains.

To further investigate the correlation between the HMP-KAP domains, the data were subjected to Pearson Correlation Coefficient. However, prior to the analysis, the data were tested for normality tests in order to make sure that the assumption was met.

Ethical Consideration

Ethical approval to conduct the study was obtained from the Human Research Ethics Committee, Universiti Sains Malaysia (Reference No: USMKK/PPP/JEPeM

[267.2 (11)]). Ethical approval was also obtained from the Malaysia Ministry of Education (Reference No: KP (BPPDP)603/5/JLD.12 (22)) since the study involved the school community.

Research Findings

Out of 301 returned instruments, only 231 (76.7 %) were usable. Almost a quarter of the instruments were discarded due to more than 10 % missing data (1 %) and also due to misfit persons (22.3 %) who did not fulfill two or more of the following criteria (Bond and Fox 2007):

- Positive Point-measure correlation (PTMEA Corr)
- The Infit and Outfit Mean Square should be between 0.5 and 1.5
- The Infit and Outfit ZSTD should be between −2.00 and 2.00
- Unidimensionality that should indicate either it is a strong (>40 %), moderate (>30 %) or minimal (>20 %) dimension (Conrad et al. n.d.).

However, those with negative PTMEA Corr were discarded from the data set even though they fulfilled the other criteria.

Demographic Profiles of the Respondents

Table 1 summarizes the demographic profiles of the school cooks in Kelantan. Majority of the respondents were from Kota Bharu district (30.7 %), followed by Tumpat (11.7 %) and Pasir Mas (11.7 %), Pasir Puteh (11.3 %), Kuala Krai (9.1 %) and Machang (9.1 %). Less than 5 % of the respondents were from Bachok, Gua Musang, and Jeli. None of the respondents represented Tanah Merah district since no response was given by the Education Officer of Tanah Merah PPD in conducting the study.

In Kelantan, majority of the school cooks were Malays (98.7 %), aged between 40 and 49 years old (39.7 %) and almost all of them were females (93.9 %). Most of them had completed their secondary school education (82.7 %). Majority the respondents had worked at the school canteen for more than 5 years (41.3 %). Most of them already hold a Food Handling certificate (94.8 %), and only 24.8 % of them hold a Healthy Catering Practice certificate. Only slightly more than half of the respondents knew that “food safety” and “nutrition” carry different definition. The remaining respondents were either “not sure” (15.8 %) or given a wrong answer (27.25) to the terms.

Table 1 Demographic profiles of the school cooks in Kelantan

Characteristic	n	%
District (n = 231)		
Bachok	17	7.4
Gua Musang	11	4.8
Jeli	10	4.3
Kota Bharu	71	30.7
Kuala Krai	21	9.1
Machang	21	9.1
Pasir Mas	27	11.7
Pasir Puteh	26	11.3
Tumpat	27	11.7
Age, years (n = 229)		
20–29	17	7.4
30–39	61	26.6
40–49	91	39.7
≥50	60	26.2
Gender (n = 231)		
Male	14	6.1
Female	217	93.9
Race (n = 231)		
Malay	228	98.7
Chinese	1	0.4
Indian	1	0.4
Others:	1	0.4
Education (n = 231)		
No education	8	3.5
Primary school	32	13.9
Secondary school	173	74.9
Certificate	11	4.8
Diploma	4	1.7
Degree	1	0.4
Others: Masters qualification	2	0.9
Foodservice experience at school canteen (n = 231), mean = 3.987 ± 1.974 year		
Less than 1 year	42	18.3
1–2 years	21	9.1
2–3 years	37	16.1
3–4 years	25	10.9
4–5 years	10	4.3
More than 5 years	95	41.3

(continued)

Table 1 (continued)

Characteristic	n	%
Hold an accredited Food Handling Certificate? (n = 231)		
Yes	219	94.8
No	12	5.2
Hold an accredited Healthy Catering Practice Certificate? (n = 230)		
Yes	57	24.8
No	173	75.2
Is "food safety" similar to "nutrition"? (n = 228)		
Yes	62	27.2
No	130	57
Not Sure	36	15.8

Classifications of the School Cooks and Their Mean HMP-KAP Scores

Table 2 represents the logit mean measure for each HMP-KAP domains while Fig. 1a, b, and c represents the variable maps for each domain representing item difficulty and person ability in responding to the questionnaire. Of the total 231 respondents, their mean knowledge, attitude and practice scores were as follows: 0.379 ± 0.758 for knowledge, 2.996 ± 1.803 for attitude, and 0.648 ± 0.652 for practice.

The school cook's HMP-KAP level was categorized based on the identified cut-off point. Prior to categorization of the school cooks, the items in each domain were sorted according to their difficulty level and later mapped to the respective cut-off point. As shown in Fig. 1 item B14 (logit 1.17) and B24 (logit 0.21) of the knowledge domain were identified to be the cut-off point for the 75 % and 50 categories, respectively while item C4 (0.48) and C3 (0.05) were identified for attitude domain. For practice domain, item D3 (logit −0.09) was identified as its 50 % cut-off point. Hence, a straight line was drawn at the respective logit to represent the proposed category (Fig. 1a, b and c). The figure shows that more than half of the respondents had moderate HMP-KAP knowledge (50.65 %) with good attitude (97 %) and practices (88.74 %) towards the healthy meal preparation. The

Table 2 Logit mean, model error, maximum and minimum logit measure according to domains of HMP-KAP questionnaire

Domain	Logit mean ± SD	Model error	Maximum	Minimum
Knowledge	0.379 ± 0.758	0.51	2.02	−1.45
Attitude	2.996 ± 1.803	0.94	5.90	−1.15
Practice	0.648 ± 0.652	0.44	2.47	−1.20

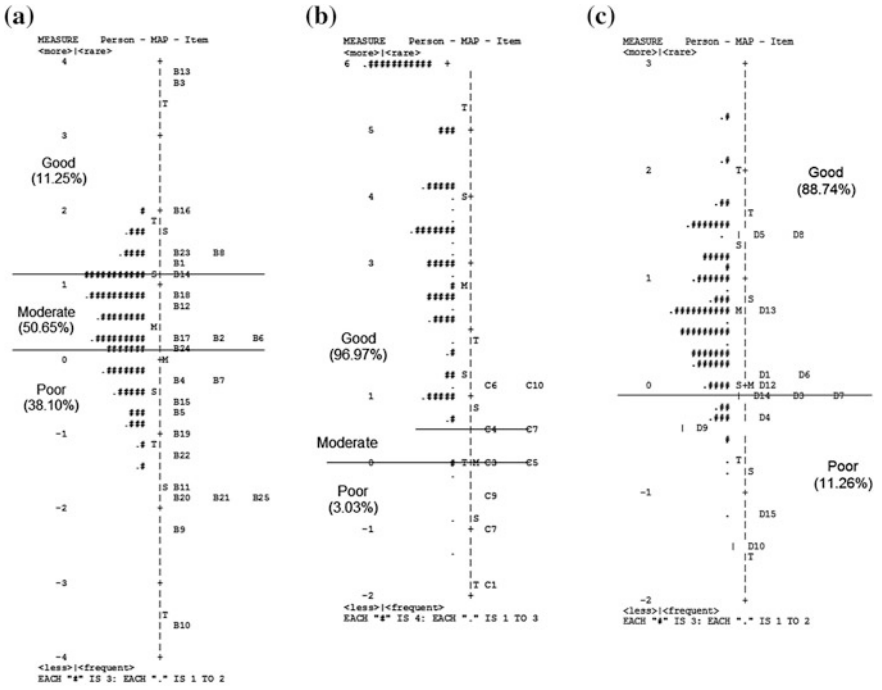


Fig. 1 Variable maps for each HMP-KAP domain reflecting categories of school cooks based on item difficulty and person ability. **a** Knowledge domain; **b** Attitude domain; **c** Practice domain

study also found that despite poor knowledge on healthy meal preparation, the school cooks' attitude and practice were excellent. Similar findings were also reported by Chen et al. (2012).

As can be seen from Fig. 1, item B13 (ways to prepare nutritious foods) and B3 (foods that would provide the most kcal content) were considered as the two most difficult items while item B9 (obesity and its related health implication) and B10 (health implication related to sugar intake) were the two easiest items to endorse. Frequent advertisement related to the implications of having too much sugar or fat foods on local television, radio, and other medium such as newspapers, magazines, and pamphlet may contribute to better knowledge related to food and health implications. However, indirect education or exposure on ways to prepare nutritious foods and to understand the food that is eaten is very scarce. Similar findings were reported by Aung et al. (2012) where the university students also had poor nutrition knowledge including to understand healthy eating concept. Poor nutrition knowledge among food handlers as well as the consumers including schoolchildren (Al-Naggar and Chen 2011; Chen et al. 2012; Mirsanjari et al. 2012) can lead to unwise food selection and finally results in chronic diseases such as diabetes, cardiovascular diseases, hypertension, and obesity.

Table 3 Summary of Spearman's rho analysis in determining the correlation between SOC and HMP-KAP domains

	Mean	SD	Knowledge	Attitude
Knowledge	0.379	0.758	–	–
Attitude	2.996	1.803	0.199*	–
Practice	0.648	0.652	0.157*	0.172*

Note $^{*}p < 0.05$, $n = 231$

Correlation Analysis of HMP-KAP Domains

Prior to any analyses, the normality of the data was assessed. The Kolmogorov–Smirnov statistic with a Lilliefors significance level for testing normally was produced with the normal probability and detrended probability plots. It was shown that the significance level for the Kolmogorov–Smirnov as well as the Shapiro–Wilks for SOC and all the HMP-KAP domains were less than $p < 0.05$, indicating that the normality could not be assumed. Hence, a nonparametric alternative to the parametric bivariate correlation (Pearson's r), that is Spearman's rho correlation coefficient was applied.

Table 3 shows the Spearman's rho summary in determining associations between the HMP-KAP domains. The analyses indicate that there were weak positive linear relationships between knowledge and attitude ($\rho = 0.199$, $n = 231$, $p = 0.002 < 0.05$), knowledge and practice ($\rho = 0.157$, $n = 231$, $p = 0.017 < 0.05$), and attitude and practice ($\rho = 0.173$, $n = 231$, $p = 0.009 < 0.05$) among the school cooks in Kelantan. The conclusion was made at the significance level, $\alpha = 0.05$ (5 %), where the value of "Sig. (1-tailed)" is less than the predetermined alpha value (0.05), thus the stated null hypothesis was failed to be accepted. The findings in this study were in agreement with Chen et al. (2012). Aung et al. (2012) however, did not find any association between knowledge score and practice.

Conclusion

Although almost 62 % of the school cooks have moderate to good HMP knowledge, none of them really understand the concept of nutritious meals, and very few were aware on foods that would provide the most kcalories. It is further proven by the fact that almost half of the respondents were either did not know or unsure that the terms "food safety" and "nutrition" do carry different definition (Table 1). It is considered as a crucial issue since school cooks are the most important people at schools to prepare and serve meals to the schoolchildren. Schools have been regarded as the best place to introduce, educate, and nurture good eating habit (Bargiota et al. 2013; De Róiste et al. 2012; Pérez-Rodrigo et al. 2001; Schwartz et al. 2011). Hence, a proactive measure such as providing continuous nutrition

education (Miller and Branscum 2012), must be done by the respective authority in improving the healthy meal preparation KAP level of the school cooks as well as other school communities including school management team, staffs and all food operators at schools. Emphasis must be given to those of lower education level where studies also revealed that those with higher qualification were shown to have moderate to good nutrition knowledge (Mohd Nasir et al. 2012; Sedek and Yih 2014) than those with lower education level. As shown in Table 1 and also reported by other studies (Abdul-Mutalib et al. 2012; Lee et al. 2012; Nee and Sani 2011), majority of the food operators at schools and other foodservice settings are those with secondary education level.

Acknowledgment The authors would like to thank our institutions, the Ministry of Higher Education of Malaysia, and the Ministry of Education of Malaysia for approving and supporting us to conduct this study. A special thank also goes to the Kelantan State Department of Education, the District Office of Education in Kelantan and schools involved for helping us in this study. The cooperation of all respondents involved is also appreciated.

References

Abdul-Mutalib, N.-A., Abdul-Rashid, M.-F., Mustafa, S., Amin-Nordin, S., Hamat, R. A., & Osman, M. (2012). Knowledge, attitude and practices regarding food hygiene and sanitation of food handlers in Kuala Pilah, Malaysia. *Food Control*, *27*(2), 289–293. http://doi.org/10.1016/j.foodcont.2012.04.001.

Abotsi, A. K. (2013). Expectations of school feeding programme: Impact on school enrolment. *Attendance and Academic Performance in Elementary Ghanaian Schools, 3*(1), 76–92.

Adelman, S. W., Gilligan, D. O., & Lehrer, K. (2008). *How effective are food for education programs? A critical assessment of the evidence from developing countries—Full text*. http://doi.org/10.2499/0896295095FPRev9.

Al-Naggar, R. A., & Chen, R. (2011). Nutrition and cancer prevention: Knowledge, attitudes and practices among young Malaysians. *Asian Pacific Journal of Cancer Prevention: APJCP, 12*(3), 691–694.

Arop, M. S., Rahman, S. A., & Fauzi, M. (2000). Evaluation of the school supplementary feeding program in Peninsular.

Arop, M. S. M., Rahman, S. A., Abdullah, A., & Jani, M. F. (2003). Nutritional status of school children receiving supplementary feeding program in Peninsular Malaysia. *Sains Malaysiana, 32*, 131–146.

Aung, P. P., Foung, C. S., Azman, K. B., Ain, N., & Zulkifeli, B. (2012). Knowledge, attitude, and practice of healthy eating among the 1st and 2nd year students of universiti Malaysia sarawak (UNIMAS). *2012 International Conference on Nutrition and Food Sciences IPCBEE*, *39*, 188–194.

Aznita Izma, M. A., Norimah, A. K., & Khairul Hasnan, A. (2009). Predicting healthy eating among normal weight schoolchildren using theory of planned behaviour. In *24th Scientific Conference of Nutrition Society of Malaysia* (p. S41).

Bargiota, A., Pelekanou, M., Tsitouras, A., & Koukoulis, G. N. (2013). Eating habits and factors affecting food choice of adolescents living in rural areas. *Hormones and Behavior*, *12*(2), 246–53. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23933693.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New Jersey: Routledge.

Chen, S., Soo, K., Azriani, A., Van Rostenberghe, H., & Sakinah, H. (2012). Nutrition knowledge, attitude and practice of teachers in rehabilitation centres in Northern Malaysia. *Mal J Nutr, 18* (2), 185–205.

Conrad, K. M., Conrad, K. J., Riley, B. B., Funk, R., & Dennis, M. L. (n.d.). Validation of the treatment satisfaction scale_likert to the Rasch measurement model, GAIN Methods Report 1. 0.

De Róiste, A., Kelly, C., Molcho, M., Gavin, A., & Nic Gabhainn, S. (2012). Is school participation good for children? Associations with health and wellbeing. *Health Education*, *112* (2), 88–104. http://doi.org/10.1108/09654281211203394.

Hamid Jan, J. M., Mitra, A. K., Hasmiza, H., Pim, C. D., Ng, L. O., & Wan Manan, W. M. (2011). Effect of gender and nutritional status on academic achievement and cognitive function among primary school children in a rural district in Malaysia. *Malaysian Journal of Nutrition, 17*(2), 189–200.

Izumi, B. T., Alaimo, K., & Hamm, M. W. (2010). Farm-to-school programs: perspectives of school food service professionals. *Journal of Nutrition Education and Behavior*, *42*(2), 83–91. http://doi.org/10.1016/j.jneb.2008.09.003.

Khatimin, N., Aziz, A. A., Zaharim, A., & Yasin, S. H. M. (2013). Development of objective standard setting using rasch measurement model in Malaysian institution of higher learning. *International Education Studies*, *6*(6), 151–160. http://doi.org/10.5539/ies.v6n6p151.

Lee, H. Y., Nadirah, W., Chik, W., Bakar, F. A., Saari, N., & Mahyudin, N. A. (2012). Sanitation practices among food handlers in a military, 2012 (November), 1561–1566. http://doi.org/10.4236/fns.2012.311204.

Miller, C. K., & Branscum, P. (2012). The effect of a recessionary economy on food choice: Implications for nutrition education. *Journal of Nutrition Education and Behavior*, *44*(2), 100–106. http://doi.org/10.1016/j.jneb.2011.01.015.

Mirsanjari, M., Muda, W. A. M. W., Ahmad, A., Othman, M. S., & Mirsanjari, M. M. (2012). Diversity of nutrient intake in pregnant women with different nutritional behaviors. In *International conference on nutrition and food sciences* (Vol. 39, pp. 153–158).

MOH (1997). A Survey of KAP on Food and Nutrition. Technical Working Group (Research), National Coordinating Committee for Food and Nutrition, Ministry of Health (MOH), Malaysia.

MOE (2011). Guidelines to Healthy School Canteen Management (1st Edition). Kuala Lumpur: Ministry of Education (MOE), Malaysia

Mohd Nasir, M. T., Norimah, A. K., Hazizi, A. S., Nurliyana, A. R., Loh, S. H., & Suraya, I. (2012). Child feeding practices, food habits, anthropometric indicators and cognitive performance among preschoolers in Peninsular Malaysia. *Appetite*, *58*(2), 525–530. http://doi.org/10.1016/j.appet.2012.01.007.

Monir, Z. M., Khalifa, A. G., Metwally, A. M., Hamid, N. A., Hanan, A., & Salah, E. M. (2013). The impact of social status and school meal on psychosocial behavior in governmental primary school children in Egypt. *Journal of Applied Sciences Research*, *9*(6), 3556–3565.

Moore, C. J., Lowe, J., Michopoulos, V., Ulam, P., Toufexis, D., Wilson, M. E., & Johnson, Z. (2013). Small changes in meal patterns lead to significant changes in total caloric intake. Effects of diet and social status on food intake in female rhesus monkeys. *Appetite*, *62*, 60–9. http://doi.org/10.1016/j.appet.2012.11.011.

Nee, S., & Sani, N. (2011). Assessment of Knowledge, Attitudes and Practices (KAP) among food handlers at residential colleges and canteen regarding food safety. *Sains Malaysiana*, *40*(4), 403–410. Retrieved from http://www.ukm.my/jsm/pdf_files/SM-PDF-40-4-2011/19Siow.pdf.

Ni Mhurchu, C., Gorton, D., Tunrley, M., Jiang, Y., Michie, J., Maddison, R., & Hattie, J. (2012). Effects of a free school breakfast programme on children's attendance, academic achievement and short-term hunger: results from a stepped-wedge, cluster randomised controlled trial. *Journal of Epidemiology & Community Health*, 6–13. http://doi.org/10.1136/jech-2012-201540.

Ni Mhurchu, C., Turley, M., Gorton, D., Jiang, Y., Michie, J., Maddison, R., & Hattie, J. (2010). Effects of a free school breakfast programme on school attendance, achievement, psychosocial

function, and nutrition: a stepped wedge cluster randomised trial. *BMC Public Health*, *10*(1), 738. http://doi.org/10.1186/1471-2458-10-738.

Nik Rosmawati, N. H., Wan Manan, W. M., Noor Izani, N. J., Nik Nurain, N. H., & Razlina, A. R. (2014). How healthy is food served at primary school canteen in Malaysia. In *6th international conference on postgraduate education*.

O'Neil, C. E., Byrd-Bredbenner, C., Hayes, D., Jana, L., Klinger, S. E., & Stephenson-Martin, S. (2014). The role of breakfast in health: definition and criteria for a quality breakfast. *Journal of the Academy of Nutrition and Dietetics*, *114*(12 Suppl), S8–S26. http://doi.org/10.1016/j.jand.2014.08.022.

Pérez-Rodrigo, C., Klepp, K.-I., Yngve, A., Sjöström, M., Stockley, L., & Aranceta, J. (2001). The school setting: an opportunity for the implementation of dietary guidelines. *Public Health Nutrition*, *4*(2b), 717–724. http://doi.org/10.1079/PHN2001162.

Schwartz, C., Scholtens, P. A. M. J., Lalanne, A., Weenen, H., & Nicklaus, S. (2011). Development of healthy eating habits early in life. Review of recent evidence and selected guidelines. *Appetite*, *57*(3), 796–807. http://doi.org/10.1016/j.appet.2011.05.316.

Sedek, R., & Yih, T. Y. (2014). Dietary habits and nutrition knowledge among athletes and non-athletes in national university of Malaysia (UKM). *Pakistan Journal of Nutrition, 13*(12), 752–759.

Vecchiarelli, S., Takayanagi, S., & Neumann, C. (2006). Students' perceptions of the impact of nutrition policies on dietary behaviors. *Journal of School Health*, *76*(10), 525–531. http://doi.org/10.1111/j.1746-1561.2006.00153.x.

Wan Nudri, W. D., & Abdullah, A. (1996). Macronutrients content in foods available in primary school canteens. *Mal J Nutr, 2*, 67–77.

Wharton, C. M., Long, M., & Schwartz, M. B. (2008). Changing nutrition standards in schools: The emerging impact on school revenue. *Journal of School Health, 78*(5), 245–251.

Science Process Skill Assessment: Teachers Practice and Competency

Norlly Mohd Isa and Hamimah Abu Naim

Introduction

Scientific knowledge is gathered and built from the both process of inquiry about natural phenomena and the content derived (Tobin and dan Capie 1980). Students use science process to learn science content (Livermore 1964). Process refers to the way science works when practised by scientists collect and interpret information that is also known as the scientific method. Science process skills always exercised in relation to some science content, and have a crucial role in the development of learning with understanding (Harlen 1999). The activities of teaching and learning based on the assessment needs to be improved in order to form meaningful information sharing to enhance the skills of students.

Science Process Skill Assessment

In Malaysia, the assessment of science process skill in a written format is introduced in the public examination besides the school-based laboratory assessment (Ministry of Education 2002). Items in this test required students to plan and design an investigation which involves the use of all the science process skills. Teachers are required to conduct school base science practical, to assess students' acquisition of the science processes. To ensure the implementation of practices and assessments carried out by implementing accurately implementing, teachers must master the concepts related to assessment, evaluation, measurement, and testing.

N.M. Isa (✉) · H.A. Naim
Faculty of Education, Universiti Teknologi Malaysia, Johor, Malaysia
e-mail: eefa_aish@yahoo.com.my

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_16

Methodology

Rasch Measurement Model is used to measure, validate, and analyze person and items relating to teachers' assessment practices and competency in Science Process Skills. The instrument is divided into three sections. Section A contains six (6) demographic items to measure the demographic factor of the respondent. Section B is to measure the Assessment Practice by the science teacher. There are 38 items of five-point Likert scale starting with scale 5 for 'Strongly Agree' and scale 1 for 'Strongly Disagree.' Section C is to measure the Science Process Skills Competency by ability of respondent to answer the 28 items correctly. The instrument was conducted among 52 science teachers of secondary school from the southern region of Malaysia. Table 1 show the item construct for each section.

The analysis used in this study are reliability and separation index, item polarity, item fit, item dimensionality, and Item-Person Map using Winsteps 3.72.3 software based on Rasch Measurement Model. The fit category measure refers to category that fits and functions as expected, meaning the higher the scale value, the more positive the measurement category values. Next the Point-Measure Correlation Coefficient (PT-Measure Correlation) is evaluated. The positive PT-Measure Correlation value shows that the items in constructs functions parallel to measure the same construct, while negative value shows the response relationship for item or person is contradicting with the variables or construct. Item fit is checked to identify the extent of items in this instrument successfully measures the intended things to be measured. The mean squared infit value and mean squared outfit for each item is

Table 1 Item construct

Section	Construct	Item	Scale
Demography		6	
Teacher's assessment competency	1. Purposes and uses of assessment 2. Analysing assessment methods 3. Effective feedback 4. Develop scoring schemes 5. Administering external assessments 6. Communicate assessment information 7. Legal and ethical	BA-7 BB-5 BC-8 BD-4 BE-4 BF-4 BG-6	5 points agreement scale
Teacher's science process skill 28 items	1. Observing 2. Classifying 3. Measuring and using numbers 4. Inferring 5. Predicting 6. Communicating 7. Using space-time	DA-4 DB-4 DC-4 DD-4 DE-4 DF-4 DG-4	Basic science process skill test

checked to identify the misfit item. The item-person map was analyzed to show the relationship between person ability with the item difficulty level (Bond and Fox 2015).

Findings and Discussion

The finding data was analyzed using Winsteps 3.72.3 software based on Rasch Measurement Model for reliability and validity test such as respondent demographic, reliability and separation index, item polarity, item fit, item dimensionality, and item-person maps.

Demographic

Respondent for this test was among 52 science teachers of secondary school from the southern region of Malaysia. There are 8 males and 44 females' teacher. 84.6 % of them are a bachelor degree holder. Most of them (94.2 %) were Science and Mathematics as their teaching options. 65.4 % have 5–15 years teaching experience. Most of them have attended the assessment training (86.5 %) and Science Process Skill training (90.4 %). Table 2 shows the demographic of the respondent.

Reliability and Separation Index

Reliability Index for a person is estimated consistency of an individual ranking on the logit scale if the respondent answers different set of item to measure the same construct (Bond and Fox 2015). Reliability Index; less than 0.60 show low reliable, 0.61–0.79 fair reliable and 0.8–1 is high reliable. Person separation index value demonstrates that the level to separate the person ability. Minimum value of separation is 2.0. If the value is less than 2.0, it means that samples are not enough to separate person ability.

From Table 3, the findings show reliability index of person is 0.92 for Assessment Practice and 0.75 for Science Process Skill Competency. The person separation index for Assessment Practise is 3.50 and Science Process Skill Competency is 1.73. The person separation index value demonstrates that there are 4 levels of person ability that can be categorized on Assessment Practise and 2 levels can be categorized on Science Process Skill Competency.

Item Reliability Index shows the difficulty variance of the items. Wide difficulty range will give high item reliability and narrow difficulty range will get the low item reliability. Low item reliability also means that sample is not enough to precisely locate the items on the latent variable. Reliability Index; less than 0.60

Table 2 Respondent demographic

Demographic	Factor	Frequency	Percentage (%)
Gender	Male	8	15.4
	Female	44	84.6
Education level	Diploma	3	5.8
	Bachelor degree	44	84.6
	Master degree	5	9.6
Teaching options	Science and mathematics	49	94.2
	Language	1	1.9
	Technical and vocational	2	3.8
Teaching experience	Less than 5 years	11	21.2
	5–10 years	22	42.3
	10–15 years	12	23.0
	15–20 years	3	5.8
	More than 20 years	4	7.7
Attended the Assessment training	Yes	45	86.5
	No	7	13.5
Attended the science Process skill training	Yes	47	90.4
	No	5	9.6

Table 3 Reliability and separation index

Section	Total items	Cronbach alpha	Reliability		Separation index	
			Person	Item	Person	Item
Assessment practice	38	0.93	0.92	0.87	3.50	2.64
SPS competency	28	0.73	0.75	0.94	1.73	3.86

show low reliable, 0.61–0.79 fair reliable, and 0.8–1 is high reliable. Item separation index value demonstrates that the level of item difficulty. Good value of item separation is 3.0 that can divide the item to high, medium, low item difficulties.

From Table 2, the findings shows reliability index of item is 0.87 for Assessment Practice and 0.94 for Science Process Skill Competency. Both are high reliable. The person separation index for Assessment Practise is 2.64 and Science Process Skill Competency is 3.86. The item separation index value demonstrates that there are 3 levels of item difficulties on Assessment Practise and 4 levels of item difficulties on Science Process Skill Competency.

Cronbach Alpha KR-20 value is an estimate of the value when persons are ordered by raw scores. It reports approximate test reliability based on the raw scores

of the sample. Cronbach Alpha KR-20 value less than 0.60 show low test reliability, between 0.61 and 0.79 show fair test reliability and 0.8–1 is high test reliability. Table 2 show the Alpha Cronbach value for Assessment Practice is 0.93, means it was high test reliability. For Science Process Skill Competency the Cronbach alpha value is 0.73, means it is in fair test reliability.

Item Polarity

Item Polarity is determined by the value of point-measure correlation (PT-Measure). If the correlation coefficient is positive, the item ability to measure the construct level is valid (Linacre 2002). Negative or 'nearly zero' values of PT-Measure correlation show the relationships for response item are contradict and not consistent with the construct (Linacre 2002; Bond and Fox 2015). The good value of PT-Measure correlation is between 0.4 and 0.8 (Linacre 2012a, b). If the value is out of this range, but the expected correlation (EXP) value is close to the PT-Measure value, the item can be acceptable.

Table 4 shows all the PT-Measure values is positive for Assessment Practice construct, means that all 38 items in Assessment Practice construct are able to measure the Assessment Practice level of respondents. However found those six items; BC35, BG36, BF26, BB02, BC03, and BC34 show the PT-measure value is less than 0.4.

Item Fit

Item fit is checked to identify the extent of items in the instrument successfully measures the intended things to be measured. The mean squared infit value and mean squared outfit for each item must be within 0.6–1.4 to be considered for rating scale measurement (Linacre and Wright 1994). The standardized (ZSTD) acceptable range is from −2.0 to +2.0 for sample size 30–300 (Bond and Fox 2015). The mean infit and outfit for person and item mean squares (MNSQ) are expected to be 1.00, and the mean standardized (ZSTD) infit and outfit are expected to be 0.0 (Linacre 2012a, b).

The findings in Table 5 show the item fit analysis for Assessment Practice. Item MNSQ infit is between 0.58 and 1.95 and the item MNSQ outfit is 0.56–2.46. The mean MNSQ infit is 0.98 and mean MNSQ outfit is 1.03. Both are close to 1.00 and can be considered good index for the item. Item ZSTD infit is between −0.9 and 3.5 and the item ZSTD outfit is from −0.9 to 4.9. The mean ZSTD infit is −0.1 and mean ZSTD outfit is 0.0. There are six items have Outfit MNSQ above 1.4 and ZSTD above 2; BC34, BC35 and BF26, BD23, BG36 and BB02. Only one item was with value of Infit MNSQ below 0.6 and ZSTD value below −2.00; which is BD10.

Table 4 Item PT-measure correlation

Assessment practice			
Item	PT-measure		
	Correlation		Exp.
BA33	*q*	0.68	0.53
BA06	*m*	0.67	0.53
BF13	*j*	0.67	0.53
BC29	*e*	0.67	0.52
BC09	*c*	0.67	0.52
BB07	*o*	0.66	0.54
BA15	*g*	0.66	0.54
BA16	*h*	0.66	0.54
BA31	*f*	0.66	0.53
BC30	*d*	0.66	0.51
BB37	*b*	0.65	0.54
BG04	*O*	0.64	0.50
BG14	*S*	0.63	0.51
BB17	*l*	0.62	0.52
BE11	*n*	0.61	0.53
BD10	*a*	0.61	0.53
BA05	*r*	0.59	0.52
BE12	*Q*	0.58	0.54
BA32	*R*	0.58	0.53
BE18	*p*	0.58	0.51
BG19	*N*	0.56	0.51
BD28	*k*	0.56	0.53
BG20	*M*	0.55	0.50
BF25	*H*	0.54	0.53
BE22	*L*	0.54	0.52
BC08	*s*	0.52	0.52
BB01	*I*	0.50	0.46
BF24	*P*	0.50	0.54
BD27	*i*	0.47	0.53
BD23	*G*	0.45	0.53
BC21	*J*	0.41	0.49
BG38	*K*	0.40	0.50
BC35	*B*	*0.39*	*0.55*
BG36	*F*	*0.38*	*0.54*
BF26	*C*	*0.36*	*0.57*
BB02	*E*	*0.36*	*0.59*
BC03	*D*	*0.33*	*0.53*
BC34	*A*	*0.27*	*0.56*

Table 5 Item fit analysis for assessment practice

Item	Measure	Model	Infit		Outfit	
		S.E.	MNSQ	ZSTD	MNSQ	ZSTD
BC34	*1.61*	*0.20*	*1.95*	*3.5*	*2.46*	*4.9*
BC35	*1.12*	*0.22*	*1.93*	*3.2*	*2.36*	*4.4*
BF26	*1.73*	*0.20*	*1.50*	*2.1*	*1.84*	*3.2*
BD23	*−0.13*	*0.28*	*1.53*	*1.9*	*1.62*	*2.0*
BG36	*0.80*	*0.24*	*1.48*	*1.8*	*1.75*	*2.6*
BB02	*2.20*	*0.18*	*1.38*	*1.8*	*1.75*	*3.1*
BF25	0.09	0.27	1.40	1.5	1.42	1.5
BC03	0.09	0.27	1.30	1.2	1.78	2.5
BB01	−1.60	0.29	1.21	1.1	1.14	0.6
BG38	−1.02	0.29	1.10	0.5	1.03	0.2
BG19	−0.61	0.29	1.07	0.4	1.04	0.2
BG20	−1.02	0.29	1.08	0.4	1.05	0.3
BE22	−0.44	0.28	1.03	0.2	1.09	0.4
BG04	−0.94	0.29	1.00	0.1	0.88	−0.3
BC21	−1.18	0.29	0.97	−0.1	1.11	0.5
BE12	0.62	0.25	0.90	−0.3	0.92	−0.3
BA32	−0.13	0.28	0.87	−0.4	0.88	−0.4
BG14	−0.61	0.29	0.86	−0.5	0.77	−0.8
BF24	0.50	0.25	0.85	−0.5	0.92	−0.2
BC08	−0.36	0.28	0.81	−0.7	0.76	−0.9
BA33	0.23	0.26	0.80	−0.8	0.71	−1.2
BB07	0.62	0.25	0.79	−0.8	0.77	−0.9
BA05	−0.53	0.28	0.80	−0.8	0.80	−0.7
BE11	0.23	0.26	0.78	−0.9	0.76	−0.9
BE18	−0.69	0.29	0.79	−0.9	0.75	−0.9
BD28	−0.13	0.28	0.76	−0.9	0.76	−0.9
BA06	0.16	0.27	0.74	−1.0	0.77	−0.9
BF13	0.02	0.27	0.75	−1.0	0.74	−1.0
BB17	−0.53	0.28	0.76	−1.0	0.72	−1.1
BA16	0.30	0.26	0.74	−1.1	0.66	−1.4
BD27	0.09	0.27	0.73	−1.1	0.74	−1.0
BA15	0.56	0.25	0.69	−1.3	0.70	−1.2
BA31	−0.13	0.28	0.67	−1.4	0.62	−1.6
BC29	−0.29	0.28	0.66	−1.5	0.62	−1.5
BC09	−0.44	0.28	0.63	−1.6	0.57	−1.8
BC30	−0.61	0.29	0.64	−1.6	0.60	−1.6
BB37	0.37	0.26	0.60	−1.8	0.60	−1.7
BD10	*0.02*	*0.27*	*0.58*	*−1.9*	*0.56*	*−1.9*
Mean	**0.00**	**0.27**	**0.98**	**−0.1**	**1.03**	**0.0**
S.D.	**0.79**	**0.03**	**0.35**	**1.4**	**0.49**	**1.7**

Science Process Skill Competency construct is a dichotomous data. Respondent will give the right or wrong answer. The mean squared infit value and mean squared outfit for each dichotomous item must be within 0.5–1.50 (Linacre 2012a, b) and the ZSTD should range from −2.0 to +2.0. Table 6 shows the finding of item fit analysis for Science Process Skill Competency construct. Item mean squared infit is between 0.72 and 1.33 and the item mean squared outfit is 0.37–1.84. Only item DD84 show the ZSTD index infit and outfit above +2.0 but the MNSQ is still in range.

Table 6 Item fit analysis for science process skill competency

Item	Measure	Model	Infit		Outfit	
		S.E.	MNSQ	ZSTD	MNSQ	ZSTD
DD70	2.26	0.41	1.33	1.2	1.84	1.5
DB85	−3.50	0.78	1.28	0.6	1.83	1.0
DD84	*0.37*	*0.30*	*1.23*	*2.3*	*1.43*	*2.0*
DD93	1.02	0.32	1.16	1.2	1.31	1.2
DC67	−1.01	0.35	1.16	0.9	1.29	1.0
DD75	−0.09	0.30	1.08	0.8	1.22	1.2
DE80	1.33	0.33	1.08	0.6	1.14	0.5
DC94	−0.09	0.30	1.07	0.7	1.11	0.6
DA77	1.12	0.32	1.03	0.3	1.10	0.4
DE90	0.73	0.31	1.02	0.2	0.96	−0.1
DE68	−1.72	0.42	1.02	0.2	0.90	−0.1
DA79	2.10	0.39	1.00	0.1	1.07	0.3
DA69	−0.78	0.33	0.99	0.0	0.93	−0.2
DB82	−2.64	0.56	0.98	0.1	1.38	0.7
DG87	1.56	0.34	0.95	−0.2	0.92	−0.1
DC78	−0.09	0.30	0.94	−0.6	0.97	−0.1
DA74	−1.41	0.38	0.93	−0.2	0.76	−0.6
DB72	−1.72	0.42	0.92	−0.2	1.34	0.9
DF71	0.00	0.30	0.92	−0.8	0.86	−0.7
DG81	3.12	0.54	0.90	−0.1	1.12	0.4
DF86	1.33	0.33	0.89	−0.7	0.99	0.1
DG91	1.56	0.34	0.88	−0.6	1.06	0.3
DE88	−0.68	0.33	0.88	−0.8	0.82	−0.7
DF83	0.00	0.30	0.86	−1.4	0.80	−1.1
DC89	−2.35	0.51	0.82	−0.4	0.77	−0.2
DF92	1.81	0.36	0.80	−1.0	0.60	−1.0
DG73	0.37	0.30	0.75	−2.8	0.68	−1.8
DB76	−2.64	0.56	0.72	−0.5	0.37	−0.9
Mean	**0.00**	**0.38**	**0.99**	**0.0**	**1.06**	**0.2**
S.D.	**1.64**	**0.11**	**0.15**	**0.9**	**0.32**	**0.9**

Item Dimensionality

The value of raw variance explained by measures should not be less than 40 % and the percentage empirical value should closely match the percentage modeled. The raw first unexplained variance contrast should not exceed 15 %, second unexplained variance contrast should not exceed the first unexplained variance contrast, third unexplained variance contrast should not exceed second unexplained contrast, and so on.

Table 7 shows the Standardized Residual Variance for Assessment Practice Construct. The value of raw variance explained by measures is 37.20 % a bit less than 40 %, but it closely matches with the modeled that is 39.20 %. The unexplained variance in first contrast is 7.70 % and it less than 15 %. All the second, third, fourth, and fifth unexplained variances did not exceed the previous unexplained variance.

Table 8 show the Standardized Residual Variance for Assessment Practice Construct. The value of raw variance explained by measures is 39.4 % a bit less than 40 % but it closely matches with the modeled that is 38.6 %. The unexplained variance in 1st contrast is 6.70 % and it less than 15 %. All the 2nd, 3rd, 4th and 5th unexplained variance is not exceed each previus unexplained variance.

Person-Item Maps

Figure 1 shows the distribution of persons' positions on the left side of the vertical line and items on the right. The items cover the range of −1.60 to +2.20 logits in

Table 7 Standardized residual variance for assessment practice construct

38 items		Empirical			Modeled (%)
Total raw variance in observations	=	60.5	100.00 %		100.00
Raw variance explained by measures	=	22.5	**37.20 %**		39.20
Raw variance explained by persons	=	10.7	17.70 %		18.60
Raw Variance explained by items	=	11.8	19.50 %		20.60
Raw unexplained variance (total)	=	38.0	62.80 %	100.00 %	60.80
Unexplained variance in first contrast	=	4.7	**7.70 %**	12.20 %	
Unexpailned variance in second contrast	=	4.3	7.10 %	11.30 %	
Unexplained variance in third contrast	=	3.4	5.60 %	9.00 %	
Unexplained variance in fourth contrast	=	3.1	5.10 %	8.10 %	
Unexplained variance in fifth contrast	=	2.5	4.20 %	6.60 %	

Table 8 Standardized residual variance for science process skill competency construct

38 items		Empirical			Modeled (%)
Total raw variance in observations	=	46.2	100.00 %		100.00
Raw variance explained by measures	=	18.2	**39.40 %**		38.60
Raw variance explained by persons	=	5.5	12.00 %		11.70
Raw Variance explained by items	=	12.7	27.40 %		26.90
Raw unexplained variance (total)	=	28.0	60.60 %	100.00 %	61.40
Unexplained variance in first contrast	=	3.1	**6.70 %**	11.00 %	
Unexplained variance in second contrast	=	2.8	6.00 %	9.90 %	
Unexplained variance in third contrast	=	2.7	5.80 %	9.50 %	
Unexplained variance in fourth contrast	=	2.2	4.80 %	7.90 %	
Unexplained variance in fifth contrast	=	2.0	4.30 %	7.20 %	

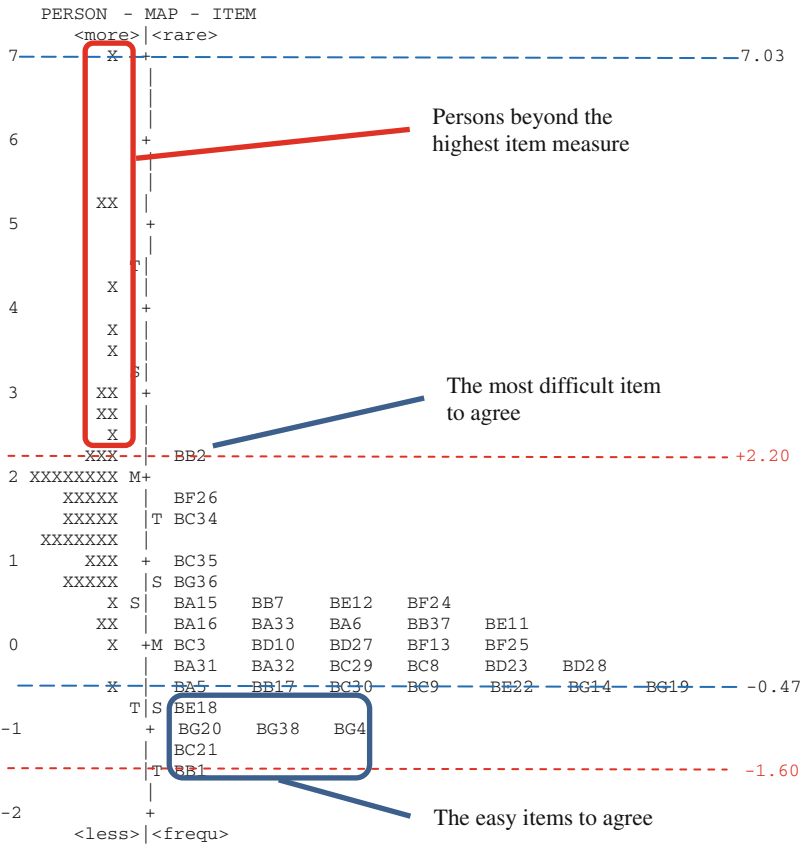


Fig. 1 Person-Item map for assessment Practice

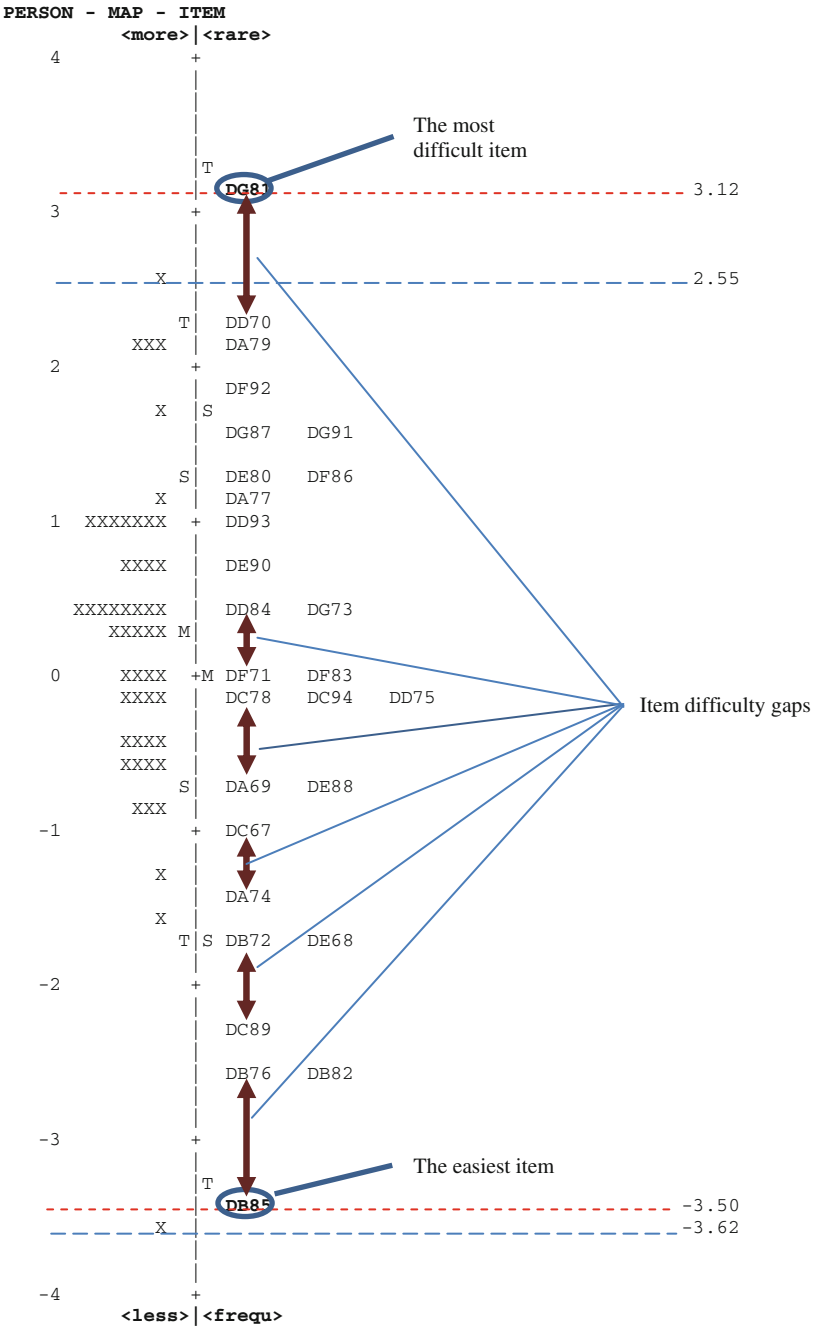


Fig. 2 Person-Item map for science process skill competency

difficulty. BB01 is the easiest item to agree and BB02 is most difficult item to agree. The highest person measure is 7.03. Person-item maps show 11 person measures is beyond the highest item measure and six item measures place below the lowest person measure.

Figure 2 shows the distribution of persons' positions on the left side of the vertical line and items on the right for Science Process Skill Competency. The items cover the range of −3.50 to +3.12 logits in difficulty. DG81 is the most difficult item and position is beyond the person's capabilities. The easiest item is DB85 but the position is still above the lowest ability person. The highest person measure is 2.55 and the lowest is −3.62. Person-item maps show the large empty space between the highest item measure and the second highest as same as the lowest to the second lowest. A person-item map also shows the gaps of the items difficulties.

Conclusion

The objective of this paper is to develop an instrument to assess the overall implementation of the practices of Science Process Skills Assessment in the classroom for subjects of science in Malaysia Secondary School. Applying Rasch analysis in instrument development is to make sure the tool use in this assessment is good and meaningful. The findings from the analysis showed that improvement is needed for some of the items to ensure that the instrument is reliable and useful. After some improvements, the test items will be distributed to sample research.

Appendix

TABLE 3.1 ASSESSMENT PRACTICE
INPUT: 52 PERSON 38 ITEM REPORTED: 52 PERSON 38 ITEM 5 CATS WINSTEPS 3.72.3

SUMMARY OF 52 MEASURED PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	150.5	38.0	1.89	.32	1.06	-.2	1.03	-.3
S.D.	12.9	.0	1.35	.07	.85	2.6	.81	2.6
MAX.	188.0	38.0	7.03	.74	4.09	7.9	3.80	8.1
MIN.	118.0	38.0	-.47	.22	.13	-4.8	.11	-5.2
REAL RMSE	.37	TRUE SD	1.30	SEPARATION	3.50	PERSON RELIABILITY	.92	
MODEL RMSE	.32	TRUE SD	1.31	SEPARATION	4.05	PERSON RELIABILITY	.94	
S.E. OF PERSON MEAN = .19								

SUMMARY OF 38 MEASURED ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	205.9	52.0	.00	.27	.98	-.1	1.03	.0
S.D.	12.8	.0	.79	.03	.35	1.4	.49	1.7
MAX.	227.0	52.0	2.20	.29	1.95	3.5	2.46	4.9
MIN.	162.0	52.0	-1.60	.18	.58	-1.9	.56	-1.9

REAL RMSE	.28	TRUE SD	.74	SEPARATION 2.64	ITEM RELIABILITY .87
MODEL RMSE	.27	TRUE SD	.74	SEPARATION 2.79	ITEM RELIABILITY .89
S.E. OF ITEM MEAN = .13					

ITEM STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
34	178	52	1.61	.20	1.95	3.5	2.46	4.9	A .27	.56	44.2	54.2	BC34
35	189	52	1.12	.22	1.93	3.2	2.36	4.4	B .39	.55	61.5	60.7	BC35
26	175	52	1.73	.20	1.50	2.1	1.84	3.2	C .36	.57	44.2	52.9	BF26
3	206	52	.09	.27	1.30	1.2	1.78	2.5	D .33	.53	65.4	71.8	BC3
2	162	52	2.20	.18	1.38	1.8	1.75	3.1	E .36	.59	34.6	49.8	BB2
36	195	52	.80	.24	1.48	1.8	1.75	2.6	F .38	.54	63.5	64.7	BG36
23	209	52	-.13	.28	1.53	1.9	1.62	2.0	G .45	.53	61.5	72.8	BD23
25	206	52	.09	.27	1.40	1.5	1.42	1.5	H .54	.53	55.8	71.8	BF25
1	227	52	-1.60	.29	1.21	1.1	1.14	.6	I .50	.46	57.7	68.6	BB1
21	222	52	-1.18	.29	.97	-.1	1.11	.5	J .41	.49	71.2	71.5	BC21
38	220	52	-1.02	.29	1.10	.5	1.03	.2	K .40	.50	76.9	72.4	BG38
22	213	52	-.44	.28	1.03	.2	1.09	.4	L .54	.52	73.1	73.6	BE22
20	220	52	-1.02	.29	1.08	.4	1.05	.3	M .55	.50	69.2	72.4	BG20
19	215	52	-.61	.29	1.07	.4	1.04	.2	N .56	.51	69.2	73.4	BG19
4	219	52	-.94	.29	1.00	.1	.88	-.3	O .64	.50	76.9	72.8	BG4
24	200	52	.50	.25	.85	-.5	.92	-.2	P .50	.54	69.2	68.5	BF24
12	198	52	.62	.25	.90	-.3	.92	-.3	Q .58	.54	71.2	66.9	BE12
32	209	52	-.13	.28	.87	-.4	.88	-.4	R .58	.53	80.8	72.8	BA32
14	215	52	-.61	.29	.86	-.5	.77	-.8	S .63	.51	76.9	73.4	BG14
8	212	52	-.36	.28	.81	-.7	.76	-.9	s .52	.52	80.8	73.5	BC8
5	214	52	-.53	.28	.80	-.8	.80	-.7	r .59	.52	75.0	73.6	BA5
33	204	52	.23	.26	.80	-.8	.71	-1.2	q .68	.53	71.2	70.8	BA33
18	216	52	-.69	.29	.79	-.9	.75	-.9	p .58	.51	76.9	73.5	BE18
7	198	52	.62	.25	.79	-.8	.77	-.9	o .66	.54	73.1	66.9	BB7
11	204	52	.23	.26	.78	-.9	.76	-.9	n .61	.53	73.1	70.8	BE11
6	205	52	.16	.27	.74	-1.0	.77	-.9	m .67	.53	75.0	71.2	BA6
17	214	52	-.53	.28	.76	-1.0	.72	-1.1	l .62	.52	75.0	73.6	BB17
28	209	52	-.13	.28	.76	-.9	.76	-.9	k .56	.53	78.8	72.8	BD28
13	207	52	.02	.27	.75	-1.0	.74	-1.0	j .67	.53	76.9	72.1	BF13
27	206	52	.09	.27	.73	-1.1	.74	-1.0	i .47	.53	78.8	71.8	BD27
16	203	52	.30	.26	.74	-1.1	.66	-1.4	h .66	.54	78.8	70.4	BA16
15	199	52	.56	.25	.69	-1.3	.70	-1.2	g .66	.54	75.0	67.9	BA15
31	209	52	-.13	.28	.67	-1.4	.62	-1.6	f .66	.53	75.0	72.8	BA31
29	211	52	-.29	.28	.66	-1.5	.62	-1.5	e .67	.52	78.8	73.5	BC29
30	215	52	-.61	.29	.64	-1.6	.60	-1.6	d .66	.51	76.9	73.4	BC30
9	213	52	-.44	.28	.63	-1.6	.57	-1.8	c .67	.52	76.9	73.6	BC9
37	202	52	.37	.26	.60	-1.8	.60	-1.7	b .65	.54	75.0	69.6	BB37
10	207	52	.02	.27	.58	-1.9	.56	-1.9	a .61	.53	84.6	72.1	BD10
MEAN	205.9	52.0	.00	.27	.98	-.1	1.03	.0			70.5	69.7	
S.D.	12.8	.0	.79	.03	.35	1.4	.49	1.7			10.8	5.8	

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
47	24	28	2.55	.60	1.33	1.0	1.21	.6	.22	.38	82.1	86.7	2 2 1 2 2 3
35	23	28	2.21	.56	.97	.0	.68	.0	.44	.42	89.3	83.7	2 2 1 4 1 2
44	23	28	2.21	.56	.87	-.4	.69	.0	.48	.42	89.3	83.7	2 2 1 3 1 2
51	23	28	2.21	.56	1.21	.8	.99	.4	.33	.42	82.1	83.7	2 2 1 6 4 3
10	21	28	1.65	.51	.88	-.4	1.45	.8	.49	.47	82.1	79.3	2 2 1 3 1 3
33	19	28	1.16	.48	.77	-1.0	1.04	.3	.59	.52	89.3	76.9	2 2 1 3 3 3
1	18	28	.94	.47	.90	-.4	.73	-.3	.59	.53	78.6	76.5	1 1 1 6 3 2
5	18	28	.94	.47	.81	-.9	.68	-.5	.63	.53	85.7	76.5	2 2 1 1 1 2
19	18	28	.94	.47	1.21	1.0	4.88	4.0	.29	.53	78.6	76.5	2 3 1 2 2 2
21	18	28	.94	.47	.99	.0	.86	-.1	.54	.53	78.6	76.5	2 2 1 3 3 2
28	18	28	.94	.47	1.41	1.8	1.70	1.2	.31	.53	64.3	76.5	1 2 1 3 3 3
43	18	28	.94	.47	1.02	.2	.81	-.2	.54	.53	71.4	76.5	2 2 1 4 3 3
45	18	28	.94	.47	.70	-1.5	.52	-.9	.69	.53	85.7	76.5	2 2 1 3 3 3
20	17	28	.71	.47	1.07	.4	.90	.0	.53	.55	71.4	76.2	2 2 1 2 2 2
24	17	28	.71	.47	1.08	.5	.93	.0	.52	.55	78.6	76.2	2 2 1 2 3
26	17	28	.71	.47	1.08	.5	.93	.0	.52	.55	78.6	76.2	2 2 1 3 3 1
27	17	28	.71	.47	1.08	.5	.93	.0	.52	.55	78.6	76.2	2 2 1 4 3 3
7	16	28	.50	.47	.82	-.8	.65	-.7	.66	.56	82.1	75.9	2 2 1 6 2 2
14	16	28	.50	.47	.99	.0	.80	-.3	.58	.56	67.9	75.9	2 2 1 4 3 2
15	16	28	.50	.47	.59	-2.2	.45	-1.4	.76	.56	89.3	75.9	2 2 1 2 2 1
16	16	28	.50	.47	.82	-.8	.65	-.7	.66	.56	82.1	75.9	2 2 1 4 2 2
17	16	28	.50	.47	1.21	1.0	1.72	1.5	.42	.56	75.0	75.9	2 2 1 3 4 3
23	16	28	.50	.47	1.10	.5	1.41	1.0	.49	.56	75.0	75.9	1 1 4 4 3 3
49	16	28	.50	.47	.99	.1	.89	-.1	.57	.56	75.0	75.9	2 2 1 4 2 2
52	16	28	.50	.47	.96	-.1	.83	-.2	.59	.56	75.0	75.9	2 2 1 4 3 3
3	15	28	.28	.47	.91	-.4	.81	-.3	.62	.57	78.6	75.6	2 3 1 3 3 2
25	15	28	.28	.47	.99	.0	1.19	.6	.55	.57	71.4	75.6	1 2 1 3 1 2
34	15	28	.28	.47	.81	-.9	.66	-.8	.67	.57	85.7	75.6	2 2 1 3 2 3
42	15	28	.28	.47	.80	-.9	.64	-.8	.68	.57	85.7	75.6	2 2 1 5 3 3
50	15	28	.28	.47	.77	-1.1	.61	-.9	.69	.57	78.6	75.6	2 2 1 3 3 3
2	14	28	.06	.47	1.01	.1	.95	.0	.58	.58	75.0	75.4	2 2 1 1 3 2
32	14	28	.06	.47	1.06	.3	.93	.0	.56	.58	67.9	75.4	2 2 1 4 3 1
38	14	28	.06	.47	.95	-.1	.83	-.3	.61	.58	75.0	75.4	2 2 1 3 2 1
48	14	28	.06	.47	.86	-.6	.66	-.8	.66	.58	67.9	75.4	2 2 1 3 3 3
13	13	28	-.15	.47	.74	-1.2	.57	-1.1	.72	.58	82.1	75.9	2 3 3 3
22	13	28	-.15	.47	1.03	.2	.96	.0	.57	.58	75.0	75.9	1 2 1 4 3 3
31	13	28	-.15	.47	1.02	.2	.92	-.1	.58	.58	82.1	75.9	2 2 1 3 3 3
36	13	28	-.15	.47	.74	-1.2	.56	-1.2	.72	.58	82.1	75.9	1 3 1 5 5 3
4	12	28	-.38	.47	.66	-1.6	.50	-1.4	.75	.59	85.7	77.1	2 2 1 1 2 2
8	12	28	-.38	.47	1.45	1.8	1.64	1.4	.36	.59	64.3	77.1	1 2 1 4 5 3
18	12	28	-.38	.47	1.30	1.2	1.14	.5	.47	.59	64.3	77.1	2 3 1 5 3 3
30	12	28	-.38	.47	.62	-1.8	.46	-1.5	.77	.59	85.7	77.1	2 2 1 2 2 3
6	11	28	-.60	.48	1.03	.2	1.20	.6	.56	.59	75.0	78.0	2 2 1 6 3 2
11	11	28	-.60	.48	.95	-.1	.90	-.1	.61	.59	82.1	78.0	2 2 1 3 1 2
29	11	28	-.60	.48	.97	.0	2.02	2.0	.56	.59	75.0	78.0	2 2 1 3 3 2
41	11	28	-.60	.48	1.07	.3	1.46	1.1	.52	.59	82.1	78.0	1 2 1 3 3 1
12	10	28	-.84	.49	.88	-.4	.81	-.3	.64	.58	85.7	79.0	2 2 1 4 2 2
39	10	28	-.84	.49	1.00	.1	1.46	1.0	.55	.58	78.6	79.0	2 2 1 3 3 3
40	10	28	-.84	.49	1.05	.3	1.50	1.1	.53	.58	78.6	79.0	2 2 1 3 2 2
9	8	28	-1.35	.52	1.86	2.6	2.31	1.8	.15	.57	60.7	81.4	2 2 1 2 1 2
37	7	28	-1.62	.54	1.34	1.1	1.33	.7	.40	.55	75.0	82.5	2 2 1 3 5 3
46	2	28	-3.62	.81	.91	.0	.59	.1	.38	.36	92.9	92.8	2 1 4 1 2 2
MEAN	14.9	28.0	.26	.49	.99	.0	1.06	.1			78.4	77.7	
S.D.	4.1	.0	1.02	.05	.23	.9	.67	1.0			7.2	3.3	

TABLE 3.1 SCIENCE PROCESS SKILL COMPETENCY
INPUT: 52 PERSON 28 ITEM REPORTED: 52 PERSON 28 ITEM 2 CATS WINSTEPS 3.72.3

SUMMARY OF 52 MEASURED PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	14.9	28.0	.26	.49	.99	.0	1.06	.1
S.D.	4.1	.0	1.02	.05	.23	.9	.67	1.0
MAX.	24.0	28.0	2.55	.81	1.86	2.6	4.88	4.0
MIN.	2.0	28.0	-3.62	.47	.59	-2.2	.45	-1.5

REAL RMSE .51 TRUE SD .88 SEPARATION 1.73 PERSON RELIABILITY .75
MODEL RMSE .49 TRUE SD .89 SEPARATION 1.82 PERSON RELIABILITY .77
S.E. OF PERSON MEAN = .14

SUMMARY OF 28 MEASURED ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	27.8	52.0	.00	.38	.99	.0	1.06	.2
S.D.	13.8	.0	1.64	.11	.15	.9	.32	.9
MAX.	50.0	52.0	3.12	.78	1.33	2.3	1.84	2.0
MIN.	4.0	52.0	-3.50	.30	.72	-2.8	.37	-1.8

REAL RMSE	.41	TRUE SD	1.59	SEPARATION	3.86	ITEM RELIABILITY .94
MODEL RMSE	.40	TRUE SD	1.60	SEPARATION	4.00	ITEM RELIABILITY .94
S.E. OF ITEM MEAN = .32						

ITEM STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
4	8	52	2.26	.41	1.33	1.2	1.84	1.5	A -.04	.30	82.7	84.9	DD70
19	50	52	-3.50	.78	1.28	.6	1.83	1.0	B .09	.28	94.2	96.2	DB85
18	25	52	.37	.30	1.23	2.3	1.43	2.0	C .13	.38	57.7	65.5	DD84
16	48	52	-2.64	.56	.98	.1	1.38	.7	D .25	.32	94.2	93.1	DB82
6	44	52	-1.72	.42	.92	-.2	1.34	.9	E .35	.35	86.5	86.0	DB72
27	18	52	1.02	.32	1.16	1.2	1.31	1.2	F .20	.36	69.2	70.4	DD93
1	39	52	-1.01	.35	1.16	.9	1.29	1.0	G .21	.37	73.1	77.5	DC67
9	30	52	-.09	.30	1.08	.8	1.22	1.2	H .28	.38	65.4	67.1	DD75
14	15	52	1.33	.33	1.08	.6	1.14	.5	I .26	.35	73.1	74.9	DE80
15	4	52	3.12	.54	.90	-.1	1.12	.4	J .29	.23	92.3	92.3	DG81
28	30	52	-.09	.30	1.07	.7	1.11	.6	K .31	.38	65.4	67.1	DC94
11	17	52	1.12	.32	1.03	.3	1.10	.4	L .31	.36	75.0	71.9	DA77
13	9	52	2.10	.39	1.00	.1	1.07	.3	M .28	.31	86.5	83.4	DA79
25	13	52	1.56	.34	.88	-.6	1.06	.3	N .41	.34	84.6	77.8	DG91
24	21	52	.73	.31	1.02	.2	.96	-.1	n .36	.37	65.4	67.2	DE90
2	44	52	-1.72	.42	1.02	.2	.90	-.1	m .36	.35	86.5	86.0	DE68
3	37	52	-.78	.33	.99	.0	.93	-.2	l .39	.37	75.0	74.3	DA69
20	15	52	1.33	.33	.89	-.7	.99	.1	k .43	.35	76.9	74.9	DF86
12	30	52	-.09	.30	.94	-.6	.97	-.1	j .43	.38	73.1	67.1	DC78
21	13	52	1.56	.34	.95	-.2	.92	-.1	i .38	.34	80.8	77.8	DG87
8	42	52	-1.41	.38	.93	-.2	.76	-.6	h .44	.36	80.8	82.5	DA74
5	29	52	.00	.30	.92	-.8	.86	-.7	g .46	.38	67.3	66.6	DF71
22	36	52	-.68	.33	.88	-.8	.82	-.7	f .49	.38	76.9	72.8	DE88
17	29	52	.00	.30	.86	-1.4	.80	-1.1	e .51	.38	67.3	66.6	DF83
23	47	52	-2.35	.51	.82	-.4	.77	-.2	d .45	.33	92.3	91.4	DC89
26	11	52	1.81	.36	.80	-1.0	.60	-1.0	c .53	.33	82.7	80.7	DF92
7	25	52	.37	.30	.75	-2.8	.68	-1.8	b .61	.38	76.9	65.5	DG73
10	48	52	-2.64	.56	.72	-.5	.37	-.9	a .57	.32	94.2	93.1	DB76
MEAN	27.8	52.0	.00	.38	.99	.0	1.06	.2			78.4	77.7	
S.D.	13.8	.0	1.64	.11	.15	.9	.32	.9			10.0	9.6	

References

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental Measurement in the human sciences* (3rd ed.). New Jersey: Lawrence Erlbaum Associates Publishers.

Harlen, W. (1999). Purposes and procedures for assessing science process skills. *Assessment in Education: Principles, Policy & Practice, 6*(1), 129.

Linacre, J. M. (2002). What do Infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2012a). Winsteps Tutorial 1, http://www.winsteps.com/a/winsteps-tutorial-1.pdf.

Linacre, J. M. (2012b). Winsteps Tutorial 2, http://www.winsteps.com/a/winsteps-tutorial-2.pdf.

Livermore, A. H. (1964). The process approach of the AAAS commission of science education. *Journal of Research in Science Teaching, 2*, 271–282.

Ministry of Education. (2002). Integrated curriculum for secondary schools. Curriculum Specifications. Science Form 1. Curriculum Development Centre, Ministry of Education.

Tobin, K. G., & dan Capie, W. (1980). Teaching process skills in the middle schools. *Journal of School Science and Mathematics, 74*(80), 590–600.

A Structural Model of Situational Constructs Accounting for Willingness to Communicate at a Japanese University

Graham George Robson

Introduction

Despite the proliferation of English through newspapers and other media in Japan, Japanese people have little interaction with English in their daily lives. In fact, it is likely some Japanese learners will never use English outside of university classes in Japan. For EFL learners like Japanese learners, therefore, the conditions of the classroom have to motivate learners to take opportunities to communicate as much as possible to develop their language skills. These conditions might include teachers creating an open and friendly atmosphere, developing learner confidence, competence, and motivation for communication in English. If the classroom conditions are right for communication, students may take opportunities to communicate, build positive experiences in the classroom that may be employed outside the classroom.

One important reason why students might or might not communicate in class is their level of willingness to communicate (WTC). Taking cues from L1 research, WTC has been used by second language researchers who have adapted the concept to L2 contexts. It has been found to be an individual difference that can support or hinder second language acquisition. If learners have higher WTC, they are more likely to communicate in the L2 (MacIntyre and Charos 1996; Yashima et al. 2004). Therefore, a better understanding of WTC can assist teachers in getting their learners to communicate more effectively in the classroom.

In Japan, too, WTC has been extensively researched, especially at the university level. A number of studies of WTC conducted in Japanese universities have used statistical models to understand which factors affect WTC (e.g., Fushino 2008; Matsubara 2011; Matsuoka 2005). Indeed, many of these researchers have moved towards explaining WTC in more dynamic, situated environments such as the

G.G. Robson (✉)
Toyo University, Hakusan Campus, Tokyo, Japan
e-mail: robson@toyo.jp

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_17

foreign language classroom, away from the L1 conception of WTC as wholly dependent on the personality of the learner. Research on WTC in the foreign language classroom can illuminate the factors that influence WTC at the point where communication takes place for these learners, namely, the classroom.

Willingness to Communicate

Whereas trait-like WTC, as described in the L1 literature, has the potential to affect people in all communication settings, there are certain situational constraints that affect what happens in a given situation. As far back as 1994, MacIntyre proposed that researchers should combine personality constructs with situational constructs to measure WTC (p. 140). This was the start of conceptualizing WTC as a situational construct for second language learners. One of the first major models of L2 WTC (MacIntyre et al. 1998), and one that would greatly influence second language research, treated WTC as both a mixture of "transient and enduring influences" (p. 546). By doing so, the researchers claimed that WTC could be used to address pedagogical concerns related to why some students speak in language classes and others do not. The authors reasoned that although many constructs can potentially affect WTC, the most dramatic change is created by the language of communication because communication in the L1 is not the same as that in the L2, given the varying degrees of competence found in L2 learners. Therefore, first and second language WTC should be conceptualized somewhat differently.

To explain what factors lead to WTC, MacIntyre et al. (1998) proposed a number of personality, affective, and situational constructs that fit into a theoretical model with six layers. Layers VI to IV represent stable and enduring factors that indirectly affect an individual's WTC. Layers I to III address factors that deal more with the specific situation of communication the learner is in. First, the Situated Antecedents (Layer III), hypothesized to include Desire to Communicate with a specific person and state communication self-confidence, are seen as the most immediate determinants of WTC. After the state constructs, the next layer is behavioral intention (Layer II), which is the intention to communicate (WTC) and the immediate antecedent of communication behavior. The authors described WTC on this level as "a readiness to enter into discourse at a particular time with a specific person or person using L2" (p. 547). Last is communication behavior (Layer I), which is the final communication event. Its position at the top of the model symbolizes the importance of the goal of learning a second language, which the authors describe as "engender[ing] in language students the willingness to seek out communication opportunities and the willingness to actually communicate in them" (p. 547). This means that by developing WTC in the classroom, students should be able to communicate both inside and, ultimately, outside the classroom. This proposed model would form the base of WTC conceptions in future L2 empirical models.

Even before MacIntyre et al. (1998), researchers had begun to look at how WTC could be applied in second language and foreign language settings contexts. Studies

in the second language setting included MacIntyre and Charos (1996) and Hashimoto (2002) and the EFL setting (Kim 2004; Yashima et al. 2004; Yu 2009). These models showed that, firstly, motivation in its various forms is a powerful predictor of WTC. Second, they found the most powerful predictors of WTC were anxiety and perceived competence, which combine to form a self-confidence factor. Third, fit indices reported in the models were acceptable. Fourth, these studies employed a measurement of trait WTC from L1 studies, and to a much lesser extent, they have not used constructs that describe what happens in the classroom.

If a difference is present between global WTC and situational WTC, it stands to reason that measures should reflect that difference (Weaver 2010). Investigating the situational nature of WTC in the East Asian context, there have been a number of empirical models that have tested constructs related to the situation, particularly at universities (Matsuoka 2005; Fushino 2008; Peng and Woodrow 2010). All these studies again used different constructs to explain WTC and reported adequate model fit indices.

Actual Communication

Looking at research investigating the relationship between WTC and actual communication behavior, as described in the MacIntyre et al. model (1998), studies have focused mainly on the intention to communicate, whether that is through trait or situational measures. Many studies have employed self-reported measures that seek to ascertain the frequency of communication learners have engaged in. This measure has often contained the same items as the trait WTC measure, but has been modified to encompass the frequency of each item (Hashimoto 2002; MacIntyre and Charos 1996). On the whole, these researchers have found a significant but low correlation between the WTC constructs and perceived frequency of communication measures (the models report correlations of up to $r = 0.33$).

What has not been researched sufficiently is whether the intention to communicate leads to actual communication in the L2. However, a few studies that address WTC measures using actual classroom behavior have found mixed results (Dörnyei and Kormos 2000; Cao and Philp 2006; Cao 2012) and have all suffered from having small sample sizes.

Main Antecedents of WTC

From the above literature review of WTC in first and second language settings, the following three antecedents were recognized: (a) Classroom Constructs, (b) Motivation Constructs, and (c) Confidence and Anxiety-Related constructs. This next section briefly covers these three parts and their relationship with WTC.

Classroom Constructs

The importance of the classroom in EFL settings needs to be recognized because what happens in the classroom has a powerful effect on learning. Within the classroom some of the possible constructs that have found to affect WTC are class-based constructs like the topic (Kang 2005), the interlocutor (Kang 2005; Weaver 2010). Also the individual effect of the teacher in the classroom (Peng and Woodrow 2010) has found to be powerful. Further, WTC was affected by the way the language was taught, namely the methodology (Aubrey 2010; Matsubara 2011) or the class sizes (Aubrey 2010). Lastly, the culture and how that affects the atmosphere of the class has been found to be important in WTC models (Peng and Woodrow 2010).

In an exemplary model by Peng and Woodrow (2010), it was found that the classroom environment was hypothesized as three subconstructs, consisting of a number of original items, namely, teacher support, student cohesiveness, and task orientation, with a combined reliability index of ($\alpha = 0.88$). The subsequent structural model found that the classroom constructs predicted learner beliefs ($\beta = 0.33$), communicative confidence ($\beta = 0.19$) and WTC ($\beta = 0.18$) and a data-driven path was added to Motivation ($\beta = 0.29$). The model fit indices were also moderate. The results of this study showed that in Asian settings the classroom can be seen as the construct that might affect all others in situational WTC models.

Motivation Constructs

Motivation in WTC studies has been conceptualized in different ways. First, the Gardner-inspired research addressed motivation as an antecedent of L2 achievement through a self-report questionnaire called the Attitude/Motivation Test Battery (AMTB; Gardner and Smythe 1981). A number of empirical and model-based studies have used motivation as described in the Gardner model and MacIntyre et al.'s (1998) conceptual model and applied them to WTC in second and foreign language settings (Hashimoto 2002; MacIntyre and Charos 1996; Yu 2009). It has been found that motivation, while not directly impacting on WTC, indirectly affects WTC through constructs such as perceived competence. These studies suffered from reliability issues related to small sample sizes and weak relationships between predicted constructs.

Second, motivation has been conceptualized as international posture (Yashima 2002) in the foreign language settings like Japan, due to the fact most people have little daily contact with foreigners. Yashima's motivational construct consisted of Interest in Foreign Affairs, Interest in Working Abroad and Cultural Friendship. Yashima's (2002) study in Japan found International Posture predicted WTC ($\beta = 0.22$) and Motivation to Learn the L2 ($\beta = 0.79$). There was no direct path from motivation to learn the L2 and WTC, but motivation led indirectly to WTC through L2 proficiency and communicative confidence. The Yashima model

has been tested successfully in different foreign language settings: South Korea (Kim 2004) and Japan (Matsuoka 2005). These studies found a clear direct link between motivation and WTC.

The final motivation conceptualization in WTC is the self-determination theory (SDT) (Ryan and Deci 2000). SDT addresses how different types of motivation are derived from the reasons or goals that bring forth a particular action. The main distinction in SDT theory is between intrinsic motivation, or goal-oriented behavior that involves "doing something because it is inherently interesting or enjoyable," and extrinsic motivation, which is goal-oriented behavior characterized by "doing something that leads to a separable outcome" (Ryan and Deci 2000, p. 55). Both of these two motivations (or orientations) can be global or context-related, but the regulation processes that underlie them are different in nature (Deci et al. 1991, p. 327). Although studies using SDT have been carried out in South East Asia, finding SDT to be greatly linked to the context, its application to WTC research is limited (Peng and Woodrow 2010; Watanabe 2011). It is clear more research is needed to confirm how SDT relates to WTC by using more context-specific instruments.

Confidence and Anxiety-Related Constructs in the L2

These constructs include anxiety, self-perceived communication competence, and confidence. These constructs, and their equivalents, have been used extensively in first and second language research. Grouped together, they have been often referred to as communication constructs as they directly affect WTC. Researchers investigating WTC models have used confidence, anxiety, and self-perceived competence in both second language (Hashimoto 2002; MacIntyre and Charos) and foreign language settings (Kim 2004; Peng and Woodrow 2010; Yashima 2002). All of these studies found higher perceived competence, lower anxiety and higher confidence to be consistent and powerful predictors of WTC. The results also showed that, generally, perceived competence is a stronger predictor than anxiety of both confidence and WTC. Finally, perceived competence and anxiety have often been measured by using the same item stems as the WTC, which came from L1 WTC research. Many studies that use models in WTC claim that WTC is situational in nature still measure perceived competence and anxiety with an instrument modified from L1 studies. More context-specific instruments are needed to measure affective constructs in situational-based research.

This brief overview of WTC antecedents reveals the need for motivation in WTC models to account for different forms of learner motivation. It also shows WTC to be heavily influenced by the situation, but only a minimal number of studies have addressed research into classroom-related constructs for WTC. It is necessary to test a wider variety of classroom-based factors measured through reliable and valid instruments. Finally, more objective data collection methods are needed that measure the nature and amount of communication taking place in the classroom.

Methods

This study reports on the results from the main study and subsequent structural model of a study of factors affecting situational WTC. Although a preliminary phase was also used in the study, it has not been reported in this study because of brevity.

Participants and Procedures

Participants come from one faculty specializing in international regional development studies at a medium-sized private university in central Tokyo. The faculty's mission statement stresses the importance of English. The preliminary instrumentation was piloted with 208 students. From there, the main study utilized a new group of students from the same faculty, which did not include study abroad students (n = 471).

In this nested design, questionnaires consisting of three parts were administered to first- and second-year student participants and one scoring rubric was given to seven English teachers (two Japanese teachers of English and five native teachers of English) who taught in the faculty in the spring semester of 2014. Participants were informed that their participation was optional. All participants agreed to answer the questionnaire.

Instrumentation

The scales were adapted and developed from previous studies. All scales were translated into Japanese by one Japanese teacher of English and back translated by another Japanese teacher of English. All items in the student survey required participants to indicate their agreement or disagreement on a 6-point Likert scale (1 = Strongly disagree; 2 = Disagree; 3 = Slightly disagree; 4 = Slightly agree; 5 = Agree; 6 = Strongly agree). The English version of the classroom and motivation subscales can be seen in Appendix A and the items covering self-perceived communication competence, communication anxiety, and willingness to communicate subscales can be seen in Appendix B.

Classroom Subscales
These subscales are based on how the classroom conditions of the classroom lead to communication, adapted from previous studies (Fushino 2008; Matsubara 2011; Peng and Woodrow 2010).

Classroom Affective Factor (CAF)
Items are based on teacher actions and perceptions, which if positive improve student perceptions of the environment for classroom communication (5 items).

Classroom Efficacy Factor (CEF)
Items are based on student perceptions of the degree to which the class is ideal for communication (6 items).

Motivation subscales
These subscales were adapted from the *Language Learning Orientation Scale—Intrinsic Motivation, Extrinsic Motivation and Motivation Subscales* (*LLOS-IEA*; Noels et al. 2003) and other SDT surveys used in foreign language settings (Otoshi and Heffernan 2011; Peng and Woodrow 2010). The stem "I speak English in class becasuse." preceded each item.

Intrinsic Motivation for Communication (IMC)
Items are based on reasons for speaking English linked pleasure gained from the activity (5 items).
Introjected Regulation for Communication (JRC)
Items are based on reasons for speaking English to save face or maintain self-worth (5 items).

External Regulation for Speaking (ERC)
Items are based on reasons for speaking English because of an external force such as tests or job hunting (4 items).

Self-perceived communication competence, communication anxiety, and willingness to communicate subscales are based on divisions including size of group (pair or whole class) (Aubrey 2010; Cao and Philp 2006), strategic competence (Canale and Swain 1980) and topics for discussion (Kang 2005).

Self-Perceived Communication Competence for whole class activities (SPC-W)
Items measure the perceived level of ability to carry out each activity in front of the whole class or speaking directly to a native teacher of English (7 items).
Self-perceived communication competence for pair and group activities (SPC-PG)
Items measure the perceived level of ability to carry out each activity in either pairs or groups (6 items).

Communicative anxiety for whole class activities (CA-W)
Items measure the perceived anxiety level when carrying out each activity in front of the whole class or speaking directly to a native teacher of English (6 items).
Communicative anxiety for pair and group activities (CA-PG)
Items measure the perceived anxiety level when carrying out each activity in either pairs or groups (6 items).

Willingness to communicate for whole class activities (WTC-W)
Items measure the willingness to communicate for each activity in front of the whole class or speaking directly to a native teacher of English (6 items).

Willingness to communicate for pair and group activities (CA-PG)
Items measure the willingness to communicate when carrying out each activity in either pairs or groups (7 items).

Composition of items for actual communication (Teacher Rubric)
Actual communication was operationalized as a product of student actions. The rubric was designed for this study to give an overall "communicative" score for each student over the duration of a 15-week semester. The final score might not reflect individual performances on specific days, but was designed to produce a mean score representative of learner communicativeness over the period. The descriptors on the rubric explained five scores as thus: 5 = highly communicative, 4 = communicative, 3 = fairly communicative, 2 = somewhat communicative, and 1 = generally not communicative.

Analysis for Main Study

This section reports on the assumptions and results leading up to the structural model.

Assumptions

The analysis was designed to (a) examine the validity and reliability of the questionnaire through both factor analysis using PASW Statistics 18.0 (SPSS Inc. 2009) and a Rasch analysis of item fit and principal component analysis (PCA) of item residuals using Winsteps (Linacre 2009). The data in the main study were analyzed acknowledging the iterative process involved in the analysis. After checking the factor analysis, it was found, first, skewness and kurtosis levels were within acceptable limits. Second, 61 univariate outliers (identified through Mahalanobis distances) and an additional 18 multivariate outliers were removed, leaving a reduced sample of $N = 471$. The removal of 79 outliers constituted 14.36 % of the total data set. Third, to investigate linearity, bivariate scatter plots were examined visually and no curvilinear relationships were identified. Fourth, there was an absence of multicollinearity and singularity, with correlations running from −0.33 to 0.75, signifying neither multicollinearity nor singularity in the data. Finally, the factorability of R was checked for each group of items by inspecting Kaiser's measure of sampling adequacy. The factorability of R ranged from 0.88 to 0.96. All these values were over the 0.60 criterion required for factor analysis (Tabachnick and Fidell 2007).

The 79 items identified and removed through the factor analysis were tentatively returned to the data set pending results of Rasch analysis. Any person with infit values beyond 2.00 was considered an outlier. In this data set, of the 550 initial respondents, 461 were retained and 89 were potential candidates for removal due to excessive infit values. Further examination revealed that many of the persons removed from the factor analysis were the same respondents that had misfit in the Rasch model.

Results

The results of the main study are summarized in Table 1. The factor analysis in this study used a generalized least squared method with a direct oblimin rotation because of the anticipated correlations between variables. The factor analysis results can be seen in the first two columns that show all factors had fair reliability and each set of subfactors accounted for a reasonable amount of variance (classroom subfactors = 52.26 %; motivation subfactors = 53.48 %; self-perceived competence subfactors = 63.10 %; communicative anxiety subfactors = 61.79 %.; willingness to communicate subfactors = 60.15 %).

The Rasch analysis was carried out in separate parts. Rasch is a statistical analysis that examines the item and person responses to make a statistical model of the data. The model produces a fit of the data to the model. The results in Table 1 for Rasch analysis are seen in the last six columns. First, analysis of the categories used in the questionnaire helps to ascertain how well participants are using the choices available in the questionnaires. Most of the subscales had to be collapsed

Table 1 Summary of subscales realized through the main study

		Factor analysis		Rasch analysis					
Group	Realized factors	Var %	Alpha	Cat	PS	IS	Var Uni %	Var Unit	Dis. att
Classroom	CAF(5)	9.55	0.80	4	2.22	15.73	61.70	1.6	0.83
	CEF(6)	42.71	0.89	5	2.36	13.30	63.83	1.6	0.86
Motivation	IMC(5)	32.76	0.89	5	2.25	12.25	62.90	1.6	0.81
	JRC(5)	14.02	0.86	5	2.48	15.12	64.12	1.6	0.84
	ERC(5)	6.70	0.81	4	2.10	17.24	71.13	1.5	0.87
Self-perc	SPC-W(7)	51.11	0.89	6	2.84	10.55	68.25	1.8	0.80
	SPC-PG(6)	11.99	0.88	5	2.17	6.39	61.20	1.8	0.83
Anxiety	CA-W(6)	13.04	0.87	4	2.14	9.57	61.35	1.5	0.87
	CA-PG(6)	48.75	0.89	6	2.53	5.82	63.14	1.5	0.88
WTC	WTC-W(6)	49.44	0.90	6	3.18	7.55	72.63	1.5	0.88
	WTC-PG(7)	10.71	0.86	5	2.24	5.66	61.34	1.5	0.82

Note Hypoth Hypothesized factors. *Var%* variance explained in factor analysis. *Alpha* reliability of construct in factor analysis. *Cat* number of categories identified in Rasch analysis. *PS* Person separation. *IS* Item Separation. *Var Uni%* shared variance of items. *Var Unit* residuals reaming after removing primary component. *Dis. Att* Disattenuated correlation. *CAF* Classroom efficacy factor; *CEF* Class efficacy factor; *IMC* Intrinsic motivation for communication; *JRC* Introjected regulation for communication; *ERC* External regulation for communication; *SPC-W* Self-perceived competence for whole class activities; *SPC-PG* Self-perceived competence for pair/group activities; *CA-W* Communication anxiety for whole class activities; *CA-PG* communication anxiety for pair/group activities; *WTC-W* Willingness to communicate for whole class activities; *WTC-PG* Willingness to communicate for pair/group activities
Numbers in brackets following the factors indicates the number of items

into either four or five choices. Next, the Rasch person and item separation estimate measure how dispersed the respondents and items are on a variable. These estimates are not bound at one, but higher values are available, depending on the differences in endorsement of items on the measure. Values of both item and person separation above 2.0 represent good separation, so all items and persons are separated reasonably in the Rasch model. After the separation, the unidimensionality was assessed through principle components analysis (PCA). The variation column signifies the common variance among the items and the following column indicates the residuals that remain after the primary component has been accounted for; this should lie below a level of 3.0 localized units (Linacre, n.d.). Finally, dimensionality is assessed by reference to the disattenuated correlation, which is the correlation of person measures resulting from items that hold either positive or negative residual loadings (that have been corrected for error). The figure should be as close to the value of one as possible, but other WTC researchers have used a benchmark of 0.80 (Elwood 2011). The results in Table 1 show that all the subfactors reached satisfactory limits through factor analysis and Rasch analysis.

Structural Equation Modeling

After the Factor and Rasch analyses, the final stage of this study was to test a structural model of classroom WTC using the EQS software (Bentler 2006). The hypothesized model is seen in Fig. 1.

The classroom situational environment is predicted to lead to motivation for communication, communication confidence, WTC and actual communication. Further, motivation for communication is predicted to lead indirectly to WTC and actual communication through communicative confidence and WTC leads to actual communication.

In the model, Fig. 1, the boxes are represented by single indicator variables from person estimates, available through the Rasch software. Using the person estimates is a way to transform the raw data into quantitative measures (Wright and Stone 1999), providing a more accurate representation of individual scores on the Likert scale.

The assumptions of this base model were checked. First, a further, 49 univariate and 26 multivariate outliers were removed, as well as 20 respondents showing higher levels of kurtosis, leaving the *n*-size at 376. The Mardia's coefficient was 8.49 and its standardized coefficient was 5.31., so both the maximum likelihood and robust solution are reported. Second, a visual examination of scatter plots revealed no curvilinear relationships. Third, multicollinearity and singularity were not found as the range of correlation between variables was −26 to 0.51, below the limit ($r > 0.90$). Fourth, the scatter plots in PASW exhibited no apparent visual sign of

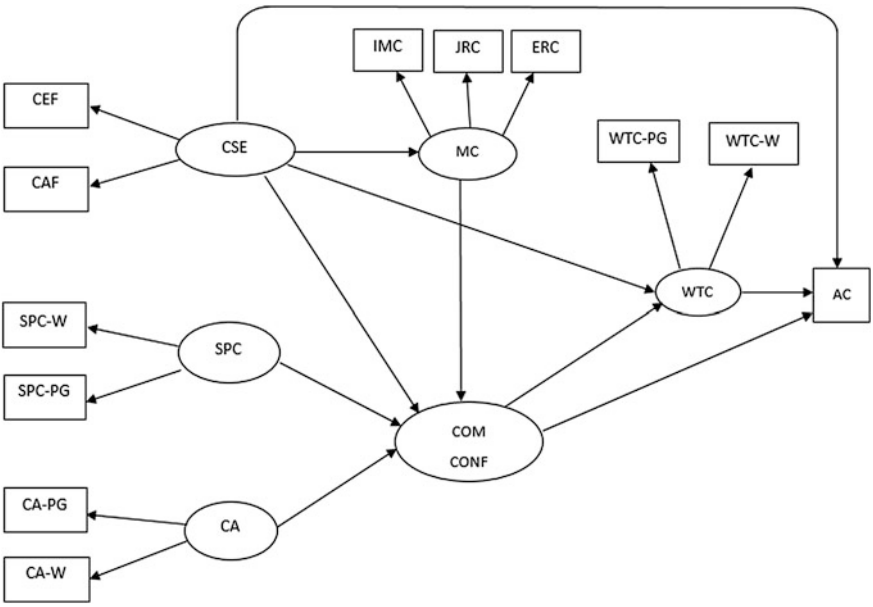


Fig. 1 Hypothesized structural model of situational WTC. *Note* Figure *CEF* Classroom efficacy factor; *CAF* Classroom affective factor; *CSE* Classroom situational environment; *IMC* Intrinsic motivation; *JRC* Introjected regulation for communication; *ERC* External regulation for communication; *MC* Motivation for communication; *SPC-W* Self-perceived competence for whole class activities; *SPC-PG* Self-perceived competence for pair/group activities; *CA-W* Communication anxiety for whole class activities; *CA-PG* Communication anxiety for pair/group activities. *WTC-W* Willingness to communicate for whole class activities; *WTC-PG* Willingness to communicate for pair/group activities. *AC* Actual Amount of classroom

heteroscedasticity. Fifth, *w*ith an *N*-size of 376 and 45 free parameters in the model, the free parameters ratio was 8.35–1. This ratio was under the rule of thumb limit suggested by Kline (2011) of 20, but much nearer to the more practical limit suggested by Kline of 10 respondents to one free parameter. Finally, the residuals were small and symmetric with 95.55 % falling in the ±0.1 range, indicating a well-specified model.

Model Estimate of the First Model and Additional Paths

The fit of the base model was quite poor. The EQS output suggested that extra variance might be explained by additional paths in the model. A number of justifiable paths were added, which can be seen in Table 2. The base model previously reported is shown in the first line.

Table 2 Modifications to base model and fit indices

Model	$ML\chi^2$	$S\text{-}B\chi^2$	$ML\Delta\chi^2$	df	CFI	SRMR	RMSEA	90 % CI	*p*
1. Base model	180.62*	156.31*		20	0.94	0.12	0.13	[0.11–0.15]	0.77
2. Revision 1 Add path MC to WTC	167.29*	146.71*	13.33*	19	0.96	0.12	0.13	[0.11–0.15]	0.80
3. Revision 2: Add path SPC to MC	109.65*	95.12*	57.64*	18	0.97	0.07	0.07	[0.05–0.10]	0.80
4. Final model: Add path CSE to SPC	33.25*	28.20*	76.40*	17	0.99	0.03	0.04	[0.00–0.06]	0.81

Note $S\text{-}B\chi^2$ Satorra-Bentler Chi-squared value; $ML\chi^2$ Maximum likelihood chi-squared value; $ML\Delta\chi^2$ Change in chi-squared value; *df* Degrees of freedom; *CFI* Comparative fit index; *SRMR* Standardized root mean square residual; *RMSEA* Root mean square error of approximation; ρ Rho reliability; $^*p < 0.001$. All $\Delta\chi^2$ Values were significant ($p < 0.001$). *MC* Motivation for communication; *WTC* Willingness to communicate; *SPC* Self-perceived competence; *CSE* Classroom situational environment

The first model modification was made by adding a theoretically justifiable path from the MC factor to the WTC factor. This was chosen as the first path because, although not part of the original model, this path has been established through regression analysis (Lu and Hsu 2008; Peng 2007) and in a structural model (Fallah 2014) related to WTC in foreign language contexts. To ascertain the degree in the amount of change from this addition, the maximum likelihood chi-squared value (with no correction for non-normality) and the Satorra-Bentler chi-squared value (assuming non-normality) were both reported. The maximum likelihood change statistic was computed via steps advised by Byrne (2006, pp. 218–219), and the Satorra-Bentler change was the difference between the chi-squared values. It was found that both of these changes were significant ($\Delta\chi^2 > 6.64$). Furthermore, the CFI and reliability improved somewhat. However, both the SRMR and RMSEA remained as they had in the first base model.

The second modification was the inclusion of a path from the SPC to MC. This path was chosen next because, although not as yet found in WTC studies, the relationship has been postulated in SDT (Ryan and Deci 2000) because competence is an antecedent of motivation. This path was reported in a study at a Japanese university (Otoshi and Heffernan 2011). After this addition, both chi-squared values fell significantly. Furthermore, the CFI rose and the SRMR, RMSEA, 90 % confidence intervals all dropped, with the reliability remaining unchanged.

The third and final modification was a path from the CSE to the SPC. This addition was chosen last because the two subfactors of classroom affective and classroom efficacy that comprise the classroom factor were new to this study. In defense of this path, Weaver (2010) reported that the nationality of the teacher or students significantly predicted Self-Perceived Competence. A cursory check of the Satorra-Bentler chi-squared value (solution assuming high multivariate kurtosis) for this final model showed a close match to the maximum likelihood solution,

meaning that the maximum likelihood is expected to be a close representation of the fit of the model. The CFI and reliability increased, and the SRMR, RMSEA and 90 % confidence intervals all decreased. All these values indicated the data was a strong fit to the model.

Final WTC Model

After removing all except one nonsignificant path, the full path model for WTC with significant paths and additional paths can be seen in Fig. 2. In addition, the correlation matrix for the 11 constructs and the single indicator variable (actual communication) from the main study are shown in Appendix C. In the final model, motivation and communicative confidence both predicted WTC, which in turn predicted actual communication. Motivation and self-perceived competence

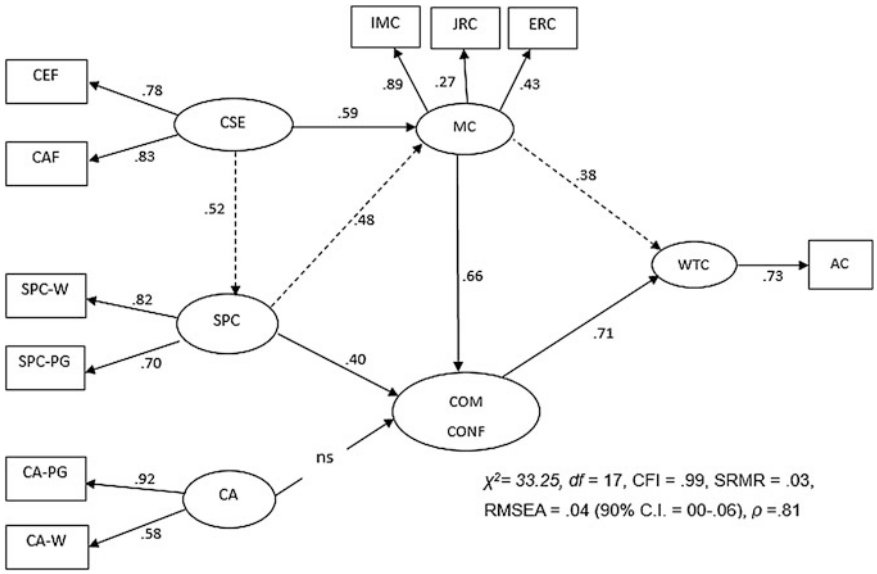


Fig. 2 Path model for the final model of WTC. *Note* Figure *CEF* Classroom efficacy factor; *CAF* Classroom AFFECTIVE FACTOR; *CSE* Classroom situational environment; *IMC* Intrinsic motivation; *JRC* Introjected regulation for communication; *ERC* External regulation for communication; *MC* Motivation for communication; *SPC-W* Self-perceived competence for whole class activities; *SPC-PG* Self-perceived competence for pair/group activities; *CA-W* Communication anxiety for whole class activities; *CA-PG* Communication anxiety for pair/group activities. *WTC-W* Willingness to communicate for whole class activities; *WTC-PG* Willingness to communicate for pair/group activities. *AC* Actual amount of classroom. *Dotted lines* indicate additional justifiable paths. *ns* nonsignificant

predicted communicative confidence. Self-perceived competence further predicted motivation. Classroom situational environment assumed a prominent position predicting both self-perceived competence and motivation. All of these paths are significant ($p < 0.01$), except the path from communicative anxiety to communication confidence. This path was retained because it was deemed important to show that communicative anxiety did not lead to communication confidence as predicted.

Discussion

In the model, the only hypothesized path found from the classroom construct was from classroom situational environment to motivation for communication. This means either the items were not appropriate, or it might be something about the context at this Japanese university, which implies that intervening factors exist between the classroom environment and other constructs. Furthermore, one additional path was added between classroom situational environment and self-perceived competence, indicating, unlike other studies, the classroom has an impact on how competent the learner feels he or she is to communicate.

As concluded in other studies, motivated students are more likely to be confident at communicating through successful experiences at communicating in the class. Along with this hypothesized result, an additional path was realized from motivation for communication to WTC, suggesting that motivation might be enough to bring a student to communicate, without attention to other affective concerns.

The path from communicative confidence to WTC was found to be significant, confirming the result in other studies, including the MacIntyre et al. (1998) WTC heuristic. WTC can be increased directly by an increase in motivation , and indirectly through increased confidence. The final additional path from self-perceived competence to motivation for communication was realized and this shows some support for the existence of psychological needs as proposed in the original SDT theory (Ryan and Deci 2000; Vallerand and Lalande 2011).

A strong path between WTC and actual communication was confirmed, a result proposed in MacIntyre et al. (1998), and finding some support elsewhere. Clearly, the intent to communicate and actual communication are two different processes.

Conclusion

This study had three limitations. First, the variance explained by the classroom situational environment and motivation for communication were a little over 50 %. Other more pertinent factors could have been added to make these factors a truer representation of their hypothesized constructs. Second, this study focused on communicative methodology, but it is impossible to ascertain what kind of English language instruction students had experienced prior to university. The final

limitation was related to the nested design of the study. Teachers in different classes might have judged students using the rubric in different ways. These all could have added bias or error into the results.

The transition from high school to university education is important in Japan. Whereas part of learners' high school classes would have been devoted to passing tests, especially the university entrance exams, university classes might be given over more to classes where students have to communicate. As the classroom is one of the main arenas for communication in the Japanese foreign language setting, it is important that conditions in the classroom are right so that factors such as motivation, confidence and, ultimately, WTC are nurtured in learners. If these factors can be attended to in the classroom through teacher action and pedagogical choices, it might lead learners to take up the opportunities to communicate outside the class.

In measuring these situational-based constructs, it is important to use situational-based instruments that actually reflect what students do in the classroom. In this study, the author attempted to identify a number of constructs that have previously been identified as important antecedents in the field of WTC and construct situational items to match those constructs. Through advanced statistical techniques like factor analysis and Rasch, it is possible to determine the reliability and validity of the factors. Finally, as in this study and other studies in WTC, the use of structural modeling is necessary to establish plausible paths that exist between these constructs in a model of WTC.

These results offer a number of pedagogical implications. First, actions taken by the teacher can clearly have an impact to reduce anxiety in the classroom. First, the teacher can manage students into pairs and groups to do activities. Group work is important because it allows students to have more speaking opportunities, potentially offering the right conditions to promote L2 acquisition (Long and Porter 1985). Groups also offer the chance for communication without the constant fear of being evaluated by the teacher. Finally, students are a lot more motivated when working in groups and discussing ideas or topics with their classmates.

The teacher's disposition can also work to lower anxiety. First, by sharing a joke or personal story with the students, the distance between teacher and students can be reduced and students are more likely to feel that the classroom is nonthreatening. Second, humor in the classroom can be a powerful tool to reduce anxiety (Matsubara 2011; Weaver 2010). A teacher that can create a fun environment is much more likely to develop learner's willingness to communicate. Third, the teacher should encourage students to experiment with language and constantly tell them that mistakes are acceptable and this may lead to language improvement.

Efficacy was the second classroom factor identified in this study. Efficacy has been described as a combination of competence and motivation (Matsuoka 2005), so working towards increasing students' feeling of competence to achieve tasks is

important. There are a couple of efforts teachers can make to raise that feeling. First, teachers can provide a pre-task phase. This usually includes planning time, either in class or before the class. By preparing ideas and suitable vocabulary to carry out a task, students increase their feelings of being able to successfully finish the task. Second, a task-based syllabus can also help to increase feelings of efficacy. Some researchers have called for implementing a task-based syllabus including tasks that increase in difficulty and complexity (Robinson 2005). Tasks can build on the skills and language acquired through previous tasks. This cycle would maintain feelings of competence if the materials are presented at a level appropriate to the learner's proficiency.

Intrinsic motivation was one of the motivation constructs highlighted in this study. According to SDT, intrinsic motivation can be brought about by satisfying the three psychological needs of relatedness, competence and autonomy. Classroom activities, like group projects, are one way to fulfill all these needs. First, projects involving other group members means that learners need to collaborate and rely on other members to complete the task. Second, as with a task syllabi, projects can build on each other so that skills are reinforced. Also, each stage of the project can offer a chance for reflection when students can assess their own performance. Finally, giving students choice over the content and how the project is carried out can help students develop feelings of autonomy.

The other main motivation subconstruct was extrinsic motivation as hypothesized through external regulation. Items in this subconstruct point towards emphasizing the need for English after graduating university. This can be achieved in a few ways. One way is to teach English that students need in the workplace. Teachers must understand the needs of students studying particular majors and design lesson plans and use authentic material that would be useful for students in the future, serving to link the classroom with the real world.

For the future, more research is needed to explore and confirm the constructs in the study. Second, more research could address the three psychological needs (autonomy, competence, and relatedness), which many SDT believe underlie intrinsic motivation. Third, more research should confirm and further explore the nature of the separation between pair and group work and whole class constructs. Fourth, more research could be carried out to incorporate the teacher as a source of assessment and data for WTC and related constructs. The teacher is able to see the students every class and obtain an impression of their abilities and individual language problems. Fifth, future research should also employ qualitative methods in combination with the numerical data. Other researchers have used mixed methods in researching WTC (Watanabe 2011; Weaver 2010). Lastly research should also focus on changes in constructs rather than the cross-sectional nature of much WTC research to date. WTC has been explored much already, but it seems there is still much to learn.

Appendix A

Classroom and Motivation Items from Main Study

Classroom affective factor (CAF)

The teacher is interested in the students
The teacher is very open to us about his/her opinions/feelings
This class provides an environment for free and open expression of ideas
The teacher sometimes talks about his/her life experiences with us
Speaking activities make me realize I need to study harder

Classroom efficacy factor (CEF)

Speaking tasks in this class are useful for me for the future
The speaking goals of this class match my own speaking goals
The teacher helps me to become a better speaker
English I learn in this class will have some benefit for me later
I can improve both my fluency and accuracy through participating in this class
Activities in this class fit the speaking needs of all students in this class

Intrinsic motivation for communication (IMC)

Speaking English is enjoyable
Speaking English well is important for me to communicate with other students/the teacher
I like to volunteer to answer questions in class and see whether my answer is correct or not
I experience a feeling of high when I speak in English
I want to develop a kind of “new me” in English

Introjected regulation for communication (JRC)

I am somehow embarrassed if I am not good at speaking English
I would be looked upon as cool if I am good at speaking English
I do not want other students to think that I cannot speak English well
Teachers will get angry if I do not speak much in class
I do not want other classmates to look down on me

External regulation for communication (JRC)

Speaking English will be useful for me in the future
I want to be completely bilingual
Improving my spoken English is necessary to help me get a better salary when I graduate
People that speak well in English in Japan are highly evaluated
If I communicate well, it might help my job hunting prospects

Appendix B

Self-perceived competence, willingness to communicate and communicative anxiety items from main study

Perceived competence	Willingness	Anxiety
I am definitely able to do these tasks/situations in my English class	I am definitely willing to do these tasks/situations in my English class	I would feel nervous doing these tasks/situations in my English class

Ask your partner/group members the meaning of an English word.
Ask your partner/group members to speak slower because you do not understand.
Talk about what you did last night to your partner/group members.
Stand in front of the class and talk about what you did last night.
Answer questions from a native teacher about a vacation you had.
Stand in front of the whole class and talk about your hobbies.
Give a self-introduction to a partner/group members.
Give a self-introduction in front of the class.
Volunteer an opinion in a whole class setting.
Do a role-play standing in front of the class in English (e.g., ordering food in a restaurant).
Ask your partner/group members how to say an English phrase to express your thoughts.
Ask your partner/group members what the time is.
Read out a two-way dialogue from the textbook in English with a partner/in a group.
Interview your partner/group members asking questions from the textbook.
Tell your teacher in English why you were late for the class.
Lead the whole class in a discussion.
Make a speech about a topic of interest to the class without using notes.
Explain how you worked out the answer to a question in the textbook to the whole class.
Answer easy questions that the teacher asks you in a whole class setting.
Give a simple agreement when your partner/other group members ask if you like a certain food.
Lead your group in a discussion.
Explain to your partner/group members how you worked out the answer to a question in the textbook.
Chat in English to your Japanese English teacher.

Appendix C

Correlation matrix for constructs from main study

Factor	1	2	4	5	6	7	8	9	10	11	12
1. CEF											
2. CAF	0.42*										
3. IMC	0.36*	0.35*									
4. JRC	−0.20*	−0.28*	−0.05								
5. ERC	−0.23*	−0.34*	−0.48*	0.25*							
6. SPC-W	0.14*	0.26*	0.35*	−0.20*	−0.31*						
7. SPC-PG	0.27*	0.28*	0.41*	−0.26*	−0.37*	0.41*					
8. CA-W	−0.24*	−0.17*	−0.29*	0.03	0.11*	−0.39*	−0.35*				
9. CA-PG	−0.29*	−0.23*	−0.31*	0.11*	0.17*	−0.27*	0.29*	0.15*			
10. WTC-W	0.25*	0.18*	0.36*	−0.17*	−0.32*	0.38*	0.41*	−0.26*	−0.29*		
11. WTC-PG	0.26*	0.26*	0.29*	−.26*	−0.35*	0.43*	0.37*	−0.31*	−0.27*	0.43*	
12. AC	0.34*	0.35*	0.41*	−0.26*	−0.23*	0.25*	0.24*	−0.38*	−0.31*	0.51*	0.66*

Note CEF Classroom efficacy factor; *CAF* Classroom affective factor; *CSE* Classroom situational environment; *IMC* Intrinsic motivation for communication; *JRC* Introjected regulation for communication; *ERC* External regulation for communication; *MC* Motivation for communication; *SPW* Self-perceived whole class; *SPP* Self-perceived pair work; *SPC* Self-perceived competence; *CAP* Communicative anxiety for pair work; *CA-W* Communicative anxiety for whole class; *Com Conf* Communication confidence; *WTW* Willingness to communicate in whole class; *WTP* Willingness to communicate in pair work; *WTC* Self-reported situational willingness to communicate; *AC* Actual amount of classroom communication. **p 0.05*

References

Aubrey, S. C. (2010). Influences on Japanese students' willingness to communicate across three different sized EFL classes. *Asian EFL Journal*. Retrieved June 20, 2012, from http://www.asian-efl-journal.com/Thesis/Thesis-Aubrey.pdf.

Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.

Byrne, B. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47. Retrieved November 28, 2014, from http://ibatefl.com/wp-content/uploads/2012/08/CLT-Canale-Swain.pdf.

Cao, Y. (2012). Willingness to communicate and communication quality in ESL classrooms. *TESL Reporter, 45*(1), 17–36.

Cao, Y., & Philp, J. (2006). Interactional context and willingness to communicate: A comparison of behavior in whole class, group and dyadic interaction. *System, 34*(4), 480–493. doi:10.1016/j.system.2006.05.002.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*(3&4), 325–346. doi:10.1080/00461520.1991.9653137.

Dörnyei, Z., & Kormos, J. (2000). The role of individual and social variables in oral task performance. *Language Teaching Research, 4*(3), 275–300. doi:10.1177/136216880000400305.

Elwood, J. (2011). *Enriching Structural Models of L2 Willingness To Communicate: The role of Personality, Ego Permeability, and Perceived Distance*. [Unpublished doctoral dissertation]. Temple University: Tokyo, Japan.

Fallah, N. (2014). Willingness to communicate in English, communication self-confidence, motivation, shyness and teacher immediacy among Iranian English-major undergraduates: A structural equation modeling approach. *Learning and Individual Differences, 30*, 140–147. doi:10.1016/j.lindif.2013.12.006.

Fushino, K. (2008). *Measuring Japanese university students' readiness for second-language group work and its relation to willingness to communicate.* [Unpublished doctoral dissertation]. Temple University: Tokyo, Japan.

Gardner, R. C., & Smythe, P. C. (1981). On the development of the attitude/motivation test battery. *Canadian Modern Language Review, 37*, 510–525.

Hashimoto, Y. (2002). Motivation and willingness to communicate as predictors of reported L2 use: The Japanese ESL context. *Second Language Studies*, *20*(2), 29–70. Retrieved September 6, 2012, from http://www.hawaii.edu/sls/uhwpesl/20(2)/Hashimoto.pdf.

Kang, S. (2005). Dynamic emergence of situational willingness to communicate in a second language. *System, 33*, 277–292. doi:10.1016/j.system.2004.10.004.

Kim, S. J. (2004). *Exploring willingness to communicate (WTC) in English among Korean EFL (English as a foreign language) students in Korea: WTC as a predictor of second language communication*. [Unpublished doctoral dissertation]. The Ohio State University, Columbus, Ohio, USA.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Linacre, J. M. (2009). *A user's guide to WINSTEPS MINISTEP: Rasch-model computer programs*. Chicago, IL: MESA Press.

Linacre, J. M. (n.d.). Dimensionality: Contrasts & variances. Retrieved December 21, 2014 from http://www.winsteps.com/winman/principalcomponents.htm.

Long, M. H., & Porter, P. A. (1985). Group work, interlanguage talk and second language acquisition. *TESOL Quarterly, 19*(2), 207–227. doi:10.2307/3586827.

Lu, Y., & Hsu, C. F. (2008). Willingness to communicate in intercultural interactions between Chinese and Americans. *Journal of Intercultural Communication Research, 37*(2), 75–88. doi:10.1080/17475750802533356.

MacIntyre, P. D., & Charos, C. (1996). Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology, 15*, 3–26. doi:10.1177/0261927X960151001.

MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *Modern Language Journal, 82*(4), 545–562. doi:10.1111/j.1540-4781.1998.tb05543.x.

Matsubara, K. (2011). *Learner attitudes towards studying English in a rural Japanese university: motivation, WTC, and preferences for instructional activities*. [Unpublished doctoral dissertation]. Temple University: Tokyo, Japan.

Matsuoka, R. (2005). *Japanese college students' willingness to communicate in English [Unpublished doctoral dissertation]*. Tokyo, Japan: Temple University.

Noels, K. A., Pelletier, L. G., Clément, R., & Vallerand, R. J. (2003). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning, 53*(1), 33–63. doi:10.1111/1467-9922.53223.

Otoshi, J., & Heffernan, N. (2011). An analysis of a hypothesized model of EFL students' motivation based on self-determination theory. *Asian EFL Journal*, *13*(3), 66–86. Retrieved May 11, 2012, from http://www.asian-efl-journal.com/PDF/September-2011.pdf.

Peng, J. E. (2007). Willingness to Communicate in an L2 and Integrative Motivation Among College Students in an intensive English language program in China. *University of Sydney Papers in TESOL, 2*, 33–59. Retrieved October 17, 2014, from http://faculty.edfac.usyd.edu.au/projects/usp_in_tesol/pdf/volume02/articl02.pdf

Peng, J. E., & Woodrow, L. (2010). Willingness to communicate in English: A model in the Chinese EFL classroom context. *Language Learning, 60*(4), 834–876. doi:10.1111/j.1467-9922.2010.00576.x.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43* (1), 1–32. doi:10.1515/iral.2005.43.1.1.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivation: Classic definitions and new directions. *Contemporary Educational Psychology, 25*, 54–67. doi:10.1006/ceps.1999.1020.

Spss Inc., (2009). *PASW Statistics for Windows, Version 18.0*. Chicago: SPSS.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Allyn and Bacon.

Vallerand, R. J., & Lalande, D. R. (2011). The MPIC: The perspective of the hierarchical model of intrinsic and extrinsic motivation. *Psychological Inquiry, 22*, 45–51. doi:10.1080/1047840x.2011.545366.

Watanabe, M. (2011). *Motivation, self-determination, and willingness to communicate by English learners at a Japanese high school.* [Unpublished doctoral dissertation]. Temple University: Tokyo, Japan.

Weaver, C. (2010). *Japanese university students' willingness to use English with different interlocutors*. [Unpublished doctoral dissertation]. Temple University: Tokyo, Japan.

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DEL: Wide Range.

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *Modern Language Journal, 86*, 54–66. doi:10.1111/1540-4781.00136.

Yashima, T., Nishide, L. Z., & Shimizu, K. (2004). The influence of attitudes and affect on willingness to communicate and second language communication. *Language Learning, 54*(1), 119–152. doi:10.1111/j.1467-9922.2004.00250.x.

Yu, M. (2009). *Willingness to communicate of foreign language learners: A WTC study in a Chinese setting*. Tallahassee, FL: VDM.

Customer Voice Retaliation (CVR) Test: Constructs Verification

Nor Irvoni Mohd Ishar and Rosmimah Mohd Roslin

Introduction

In the past, studies on customer complaining behavior have received a considerable attention in the marketing literature. However, despite the extensive research conducted, our understanding on its overall concept is still scarce, over simplified and does not reflect the full spectrum of the subject. Indeed, past research conducted on customer reaction from dissatisfied consumption experience have only focused on certain aspects of behavioural responses such as switch and complain or commonly known as complaining behavior (Singh 1988). It overlooks other possible aggressive response behaviours that might be performed by customers such as retaliation (Funches et al. 2009; Huefner and Hunt 2000). Therefore, the aim of this study is to explore customer retaliation as an extension to customer complaining behavior.

To help facilitate this study, a framework was developed based on pervious literatures to measure customer retaliation. The theory of equity is employed as the underpinning theory to explain the big picture of customer retaliatory behavior from dis-satisfied experience. According to Equity Theory (ET), people value fair treatment and have their own perception of fairness that serves as the basis to develop beliefs about what is a fair exchange reward (Adams 1963). Equity Theory also contended that when a person feels that the system or process is unjust, they will make attempts to achieve fairness or equitable relationship (Adams 1963; Pritchard 1969), or in other words they retaliate. There are many different ways that a customer can retaliate to achieve fairness. One of it is by voicing dissatisfaction aggressively. Therefore, for this study the term customer voice retaliation (CVR) will be used. However, in order for voice retaliation to take place, it requires

N.I. Mohd Ishar (✉) · R. Mohd Roslin
Faculty of Business and Management, Universiti Teknologi MARA, Shah Alam, Malaysia
e-mail: irvoni@gmail.com; irvoni@salam.uitm.edu.my

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_18

a strong trigger. Often times the trigger is in the form of dissatisfied experience, and emotion. Therefore, in this study we will be investigating customer's dissatisfied experience attribution (DExSA) and emotional experience (EMEx) in relation to CVR.

Traditionally, researchers have been applying the Classical Test Theory (CTT) to analyze data. This scenario however has changed. Although many testing and measurement textbooks coined CTT as the only way to determine the quality of an assessment, the Item Response Theory (IRT) such as Rasch Measurement Model, does offer a sound alternative to the classical approach (Idowu et al. 2011). Indeed, Rasch Analysis has gained considerable attention among the social scientists and has been applied in various area of studies (De Battisti et al. 2005; Brentari and Golia 2008; Nor Irvoni and Mohd Saidfudin 2012; Nor Irvoni and Rosmimah 2016).

Rasch Analysis for Constructs Verification

Conventionally, social scientists rely on EFA to assess dimensionality, preliminary validity, and reliability aspects of their research instrument (Yau et al. 2007; Yoon et al. 2010). However, to ascertain the psychometric properties of an instrument, it demands for rigor and robustness in the methodological approach. It should not be restricted to domain specification but instead should aim at covering a range of the construct as wide as possible (Salzberger 2000). On that note, Rasch is seen as more appropriate method of analysis for construct verification process. Rasch represents a different philosophy of construct operationalization, provides superior foundation for assessing content validity as well as construct validity (Salzberger 2000).

Indeed, Rasch equips construct validation process with more rigor and robust analysis as it emphasizes for the items to cover different intensity level so that the entire breadth of the construct is represented (Ganglmair and Lawson 2003). Another important property of Rasch is its ability to provide researchers with interval level data. Although research using CTT has been assuming that the Likert-scale is interval, often times the response categories have a rank order, in which the intervals between the value labels cannot be presumed as equal (Jamieson 2004). Hence, the application of Rasch Rating Scale Model is most appropriate as it transforms the counts of endorsement in each response category of the rating scale into an interval scale (known as logits) based on the actual data (Grodin and Blais 2010). As a result, measures will be more meaningful, and the features of validity in terms of interpreting measures especially construct and content validity can be investigated within the Rasch measurement framework (Abdullah and Lim 2013; Smith 2005).

Rasch analysis is a method to obtain measures which are objective, fundamental, and linear and has been widely applied in education related research. In education, it is used to separate the ability of respondents and the quality of a test. It predicts the likelihood of how a person of different ability level for a particular trait should

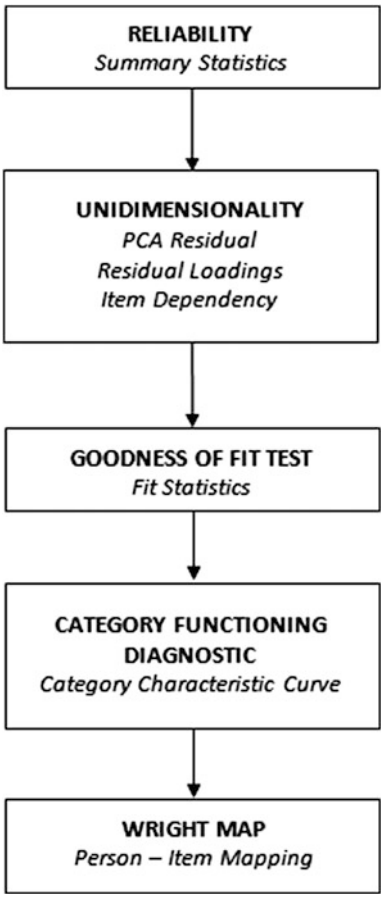


Fig. 1 CVR construct verification process

respond to an item of a certain level of difficulty. The probability of success depends on the difference between the ability of the person and the difficulty of the item (Bond and Fox 2015). From a marketing context, the term difficulty can be replaced by how hard it is to endorse an item or how extreme the item is (Ganglmair and Lawson 2003). Therefore, for CVR constructs verification, we followed the process as illustrated in Fig. 1.

Methodology

A. The Instrument

The instrument for this study comprises of three constructs represented by 66 items adapted from various sources where items were modified to suit the local context of

the study. Six point Likert rating scales were used across all three constructs. Out of 66 items, 30 items are related to dissatisfied service experience attribution (DExSA), 16 items are on negative emotional experience (NEMEx), and 20 items are on customer voice retaliation (CVR). None of the items are negatively worded.

B. Pre-test Study

Prior to pilot testing, the instrument has undergone a pre-test for content and face validation. The pre-test was conducted in two phases. The first phase involves three domain area experts, three industry practitioners, and three Rasch Measurement practitioners. Feedbacks from experts and practitioners were used to revise unclear terms, and poorly worded questions. The second stage of the pre-test involves distributing the questionnaire to 27 subscribers in three separate sessions. From the pre-test, although alpha correlation score was high (0.96) and none of the Point Measure Correlation (PMC) register a negative value; the instrument did not fulfill the unidimensionality test of Rasch Analysis.

Further, although, the variance in data explained by measures is at 66.2 %, the eigenvalue unit is at 10.9, suggesting the existence of a second dimension with the strength of 10 items and hence need further examination. Other than that, there were also feedbacks from the respondents on certain terminology used in the questionnaire. Terms such as 'blow the whistle' and 'denigrated' were highlighted as unfamiliar. These, feedbacks were taken into consideration and the wordings were changed accordingly to suit the local context. A language expert from the International Islamic University of Malaysia was consulted and necessary corrections were made before the questionnaire was released for the pilot study.

C. Pilot Study

In identifying the suitable respondents for the study, the researchers applied the mall intercept technique. At selected locations, passers-by were asked if they would like to participate in the survey. Interested participants were then asked two qualifying question prior to answering the survey. The questions are; (i) have they experienced any dissatisfaction with their mobile telco service provider, and (ii) have they shared or spoken to anyone about their dissatisfaction. Only those who answered 'yes' to both questions qualified to answer the survey questionnaire. For this pilot test, a total of 66 mobile telco subscribers fulfilled both criterion, and were handed a set of questionnaire to fill in. They were given approximately 15–20 min and survey instrument were collected immediately after the completion time. From a total of 66 questionnaires, only 53 were Rasch analyzed using Winsteps 3.80.1. The other 13 were excluded due to straight lining response pattern.

Results and Discussions

A. Reliability

Summary statistics on 66 items and 53 persons are tabulated in Table 1. The Cronbach Alpha (KR-20) Person Raw Score Test Reliability was used to test for the internal consistency of the respondents' responses and can be considered a perfectly adequate index of the interitem consistency reliability (Sekaran 2003). For this study, it was found that the Cronbach Alpha value for the three constructs is considered as 'good' (Fisher 2007) with the value of 0.90, 0.94 and 0.91 for DExSA, EMEx, and CVR respectively. This is an indication that instrument has good internal consistency in measuring the latent traits.

Further, to ensure that the person fit the Rasch model reasonably well, the data need to fulfill the fit test conditions. From Table 1, results indicate that both the person Outfit Mean square, and z-standard values are very close to the expected value of '1' and '0' which is to be expected at the norm. Hence, it can be said that respondents for this pilot test do fit the Rasch model. Other than that, from the separation index, it can be concluded that the instrument is able to reliably separate the respondents apart into 3 to 4 distinct groups. In addition, person reliability index indicates that this order of item hierarchy will be replicated with a high degree of probability if the items were given to other comparable cohorts.

Table 1 Summary statistics (Items and person reliability coefficients)

		DExSA	EMEx	CVR
Items	Reliability	0.95	0.94	0.96
	Outfit MnSq	1.02	1.0	1.0
	Outfit Z-std	0.0	−0.1	−0.1
	Separation	4.54	3.82	4.81
	Max measure	1.58	2.00	1.23
	Min measure	−1.00	−0.87	−1.32
	Model error (Mean)	0.15	0.15	0.15
Persons	Cronbach alpha (KR-20)	0.90	0.94	0.91
	Reliability	0.88	0.93	0.88
	Outfit MnSq	1.02	1.0	1.0
	Outfit Z-std	−0.1	−0.2	0.0
	Separation	2.73	3.67	2.76
	Max measure	1.02	3.47	1.57
	Min measure	−2.06	−1.99	−2.37
	Model error (Mean)	0.21	0.28	0.24

Table 2 Principle component analysis

Constructs variance explained	DExSA			EMEx			CVR		
	Empirical		Model	Empirical		Model	Empirical		Model
Total raw variance in observations	54.2	100	100	38.9	100	100	43.4	100	100
Raw variance explained by measures	24.2	44.6	44.6	22.9	58.9	58.8	23.4	53.9	55.2
Raw variance explained by persons	4.2	7.7	7.7	9.9	25.4	25.4	5.8	13.4	13.7
Raw variance explained by items	20.0	36.9	36.9	13.0	33.4	33.4	17.5	40.5	41.4
Raw unexplained variance (total)	30.0	55.4	55.4	16.0	41.1	41.2	20.0	46.1	44.8
Unexplained variance in 1st contrast	4.3	8.0	14.4	3.1	7.9	19.3	4.6	10.6	23.1

B. Unidimensionality

An important part of Rasch validity analysis is unidimensionality. It is based on the value of raw variance explained by measure and unexplained variance in 1st contrast produced by Principle Component Analysis (PCA). The results of PCA for all 3 constructs are tabulated in Table 2.

From Table 2, raw variance explained by measures for all three constructs register a value of more than 40 %, and are nearly identical to the variance expected by the model, suggesting a strong principal measurement dimension (Conrad et al. 2011). Meanwhile, unexplained variance in 1st contrast showed an acceptable fair percentage because it was less than 15 % (Fisher 2007). Both percentages of raw variance explained by measures and unexplained variance in 1st contrast indicated that the 66 items-instrument used for measuring all three construct achieved the good criteria as it met the unidimensionality trait and was able to measure what it was intended to measure. As such, the analysis showed that the data for the 66 items had a very good fit to the Rasch measurement model and supports unidimensionality. Indeed, the PCA of the Rasch Model residual indicated that the underlying items for each constructs in the instrument are assessing a unidimensional measurement model.

Table 3 highlights the items that are suspected as problematic in which might contribute to a secondary dimension. Problematic items are, items with high residual loadings value (contrast loading >+0.6 and <−0.6). Overall, 10 out of 66 items did not fulfill the loading criteria. Items are COR_3 and SOC_3 from DExSA, rag_1 and rag_2 from EMEx, and PAt_4, WOM_3, WOM_4, VIN_2, VIN_4 and 3P_4 from CVR. Therefore, there is a need to cross check if any of these items are listed in the misfitting item list that violate the goodness of fit conditions before excluding them from the final survey.

Table 3 Standardized residual loadings for items

Constructs	Loading	Measure	MnSq		Item—Domain
			Infit	Outfit	
DExSA	0.72	0.56	0.98	1.00	20 CORS_3
	0.70	0.95	0.74	0.75	27 SOC_3
EMEx	0.67	0.38	0.65	0.64	5 rag_1
	0.62	0.22	0.88	0.84	6 rag_2
CVR	0.71	−0.79	0.59	0.59	20 PAt_4
	0.65	−1.32	0.85	0.86	3 WOM_3
	0.64	−0.47	0.69	0.68	4 WOM_4
	−0.68	0.39	0.95	0.88	6 VIN_2
	−0.61	0.32	1.07	0.97	8 VIN_4
	−0.61	0.44	1.35	1.27	20 3P_4

Rasch analysis has a mean of showing redundancy of items in which can be an indication for item deletion. Indeed, redundancy enables items reduction in order to shorten the length of an instrument (Green and Frantom 2002). Raw score residual correlations are used to detect dependency between pairs from the same domain. Items that are highly locally dependent (correlation >+0.7) share more than half of their "random" variance, suggesting that only one of the pair is needed for measurement. Hence, pairs form the same domain that have Large Standardised Residual Correlations are candidates for deletion. In this study, none of the items in Table 4 violates the local dependency criteria. Hence, none will be considered for deletion.

Another form of redundancy can be detected from items that have "same measure and same domain". In Table 5, is a list of items from each constructs by measures. Any two or more items that have the same measure and also testing on the same domain are not allowed to co-exist as they are measuring the same thing at the same difficulty level. To avoid the redundancy, item that is of lower quality (with negative PMC values) needs to be eliminated from the instrument. Scrutiny of items from the same dimension having the same measure indicates that although there are items in the respective constructs having the same measures (DExSA: 7 & 30, 11 & 28, 23 & 10; EMEx: 10 & 15; CVR: 1 & 21), none are measuring in the same domain indicating no redundancy.

C. Goodness of Fit

In assessing goodness of fit, Rasch requires the items to satisfy all three important fit index conditions. The results of fit indices for suspected problematic items for each constructs are tabulated in Table 6.

Table 4 Item dependency according to largest standardized residual correlations

Construct	Correlation	Item—Domain	Item—Domain
DExSA	0.68	12 PRO_1	13 PRO_2
	0.65	7 CON_2	30 SOC_6
	0.64	14 PRO_3	17 PRO_6
	0.59	25 SOC_1	10 CON_5
	0.55	1 eSE_1	26 SOC_2
	0.54	24 CORS_7	9 CON_4
	0.52	20 COR_3	29 SOC_5
	0.49	20 COR_3	27 SOC_3
	0.44	21 COR_4	27 SOC_3
EMEx	0.48	3 ang_3	9 dis_1
	0.43	5 rag_1	7 rag_3
	0.42	5 rag_1	6 rag_2
	0.40	15 sad_3	16 sad_4
	0.39	6 rag_2	7 rag_3
	0.37	2 ang_2	13 sad_1
	0.36	1 ang_1	2 ang_2
	−0.44	5 rag_1	12 dis_4
	−0.42	6 rag_2	14 sad_2
	−0.41	2 ang_2	7 rag_3
CVR	0.68	7 VIN_3	12 COL_1
	0.67	17 3P_1	20 3P_4
	0.66	6 VIN_2	8 VIN_4
	0.59	18 3P_2	20 3P_4
	0.58	7 VIN_3	14 COL_4
	0.52	4 WOM_4	22 PAt_2
	0.51	7 VIN_3	16 COL_4
	0.50	5 VIN_1	18 3P_2
	−0.53	3 WOM_3	6 VIN_2
	−0.50	5 VIN_1	22 PAT_2

One of the first things to observe in the data is the value of Point Measure Correlation (PMC). For this pilot study, all 66 items in the instrument have a positive PMC indicating that the instrument is measuring in the right direction. However, although the responses are moving in the same direction, 7 items (see Table 6) were identified as needing closer investigation as it violated the MnSq ($0.5 > y > 1.5$), and z-std ($-2 > z > +2$) conditions. An item is a misfit when its' MnSq and z-standard values are not within the stipulated acceptable range.

Table 5 Item dependency by measure order

DExSA			EMEx			CVR		
Item no	Measure	Domain	Item no	Measure	Item	Item no	Measure	Domain
15	1.58	PRO_4	8	2.00	rag_4	15	1.23	COL_3
21	1.43	COR_4	7	0.49	rag_3	7	0.84	VIN_3
14	1.11	PRO_3	5	0.38	rag_1	19	0.79	3P_3
19	1.01	COR_2	14	0.31	sad_2	16	0.71	COL_4
27	0.95	SOC_3	6	0.22	rag_2	14	0.63	COL_2
17	0.92	PRO_6	11	−0.01	dis_3	18	0.60	3P_2
29	0.69	SOC_5	13	−0.03	sad_1	13	0.56	COL_1
20	0.56	COR_3	10	−0.05	dis_2	20	0.44	3P_4
5	0.50	eSE_5	15	−0.05	sad_3	5	0.41	VIN_1
18	0.42	COR_1	16	−0.18	sad_4	6	0.39	VIN_2
7	0.40	CON_2	1	−0.23	ang_1	8	0.32	VIN_4
30	0.40	SOC_6	4	−0.29	ang_4	17	0.24	3P_1
8	0.05	CON_3	3	−0.48	ang_3	4	−0.47	WOM_4
26	0.03	SOC_2	2	−0.60	ang_2	2	−0.76	WOM_2
22	−0.02	COR_5	9	−0.62	dis_1	22	−0.78	PAt_2
1	−0.09	eSE_1	12	−0.87	dis_4	24	−0.79	PAt_4
13	−0.46	PRO_2	–	–	–	23	−0.83	PAt_3
2	−0.48	eSE_2	–	–	–	1	−1.11	WOM_1
11	−0.53	CON_6	–	–	–	21	−1.11	PAt_1
28	−0.53	SOC_4	–	–	–	3	−1.32	WOM_3
24	−0.61	COR_7	–	–	–	–	–	–
3	−0.63	eSE_3	–	–	–	–	–	–
6	−0.67	CON_1	–	–	–	–	–	–
9	−0.69	CON_4	–	–	–	–	–	–
12	−0.73	PRO_1	–	–	–	–	–	–
4	−0.85	eSE_4	–	–	–	–	–	–
25	−0.90	SOC_1	–	–	–	–	–	–
23	−0.94	COR_6	–	–	–	–	–	–
10	−0.94	CON_5	–	–	–	–	–	–
16	−1	PRO_5	–	–	–	–	–	–

From Table 6, only 1 item (sad_2) registers MnSq value that is outside the stipulated range (Infit MnSq: 1.66 logits/Outfit MnSq: 1.59 logits). Meanwhile, for z-std, 4 items (PRO_4, PRO_3, sad_2, and PAt_3) were found to not satisfying the acceptable range of −/+ 2 logits. Therefore, all these items will be checked against the residual loadings, and item dependency values. If the same item appears as problematic, the items are a candidate for deletion and will not be included in the actual study. However, if the items were not identified as problematic, the items can be considered as fit and will be retained for actual survey.

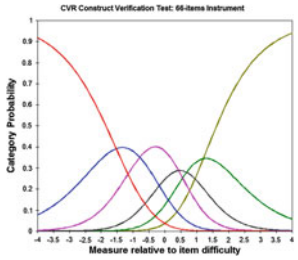
Table 6 Misfitting items by construct

Construct	Measure	Infit		Outfit		PMC	Item—Domain
		MnSq	z-STD	MnSq	z-STD		
DExSA	1.58	1.53	2.4	1.50	2.3	0.38	15 PRO_4
	1.11	1.50	2.3	1.51	2.4	0.37	14 PRO_3
EMEx	0.31	1.66	2.9	1.59	2.7	0.53	14 sad_2
	−0.87	1.29	1.5	1.56	2.4	0.60	12 dis_4
CVR	−0.83	1.42	2.1	1.61	2.9	0.07	7 PAt_3
	0.63	1.48	2.1	1.29	1.3	0.54	14 COL_2
	0.56	1.47	2.1	1.33	1.5	0.47	13 COL_1

D. Category Functioning Diagnostic

The final step of our construct verification process is rating scale calibration. It is a process by which the categories used in the instrument were analyzed for its functionality. This process is very crucial in any measurement because its validity significantly affects measurement precision. Indeed, rating scale is the way how researchers communicate with respondents as they attempt to use the response restrictions (Bond and Fox 2015). However, this process it is often overlooked.

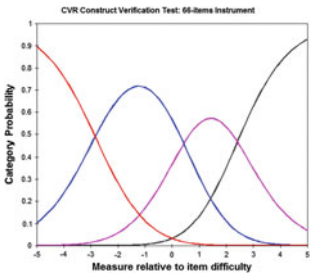
A valid scale is when all categories are functioning optimally in which enough data are represented in each thresholds. The difference in the threshold should be 1.4 logits apart but not exceeding 5 logits (Bond and Fox 2015; Linacre 1999). Rasch rating scale model (RSM) has the capacity to provide evidence for such claims as it allows researcher to extract the most meaning from the data. However, if the categories are not functioning as expected, then collapsing will take place as remedy. Therefore, the 6 response categories used in this pilot study will be re-examined to determine which categorization of responses yielded higher-quality measures. Indeed, revision of the rating scale should be done at the pilot phase in the development of the measure (Bond and Fox 2015), prior to actual data collection. Figure 2 depicts the probability curves for CVR construct.



SUMMARY OF CATEGORY STRUCTURE. Model="R"

Cat Label	Observed Count	Average Measure	Infit MNSQ	Outfit MNSQ	Threshold	Category Measure	
1	537	-1.42	.87	.91	NONE	(-2.89)	1
2	830	-.79	1.00	.99	-1.53	-1.33	2
3	1007	-.34	1.04	1.02	-.79	-.30	3
4	563	.05	1.05	1.10	.43	.48	4
5	360	.37	1.09	1.09	.68	1.29	5
6	198	.81	.98	1.05	1.20	(2.63)	6
MISSING	3	0	-.11				

Fig. 2 Six-rating scale category characteristic curve (CCC)



SUMMARY OF CATEGORY STRUCTURE. Model="R"

Cat LABEL	Observed Count	OBSVD Measure	Infit MNSQ	Outfit MNSQ	Threshold	Category Measure	
1	537	-2.38	.93	.94	NONE	(-3.97)	1
2	1837	-.85	.99	.97	-2.84	-1.23	2
3	923	.30	1.05	1.06	.43	1.43	4
4	198	1.39	1.01	1.04	2.41	(3.60)	6
MISSING	3	0 -.24					

Fig. 3 Four-rating scale category characteristic curve (CCC)

Analysis revealed that the original six-rating scale does not function effectively as depicted in Fig. 2. Overlapping peaks indicates that respondents are not able to differentiate between the categories in which will disrupt construct definition. This is further supported by the threshold estimates values between the categories which are less than 1.4 logits. Therefore, collapsing problematic categories would be a good remedy to overcome such problem. There are two general rules for collapsing categories. First it has to be logical, and second it should create a more uniform frequency distribution (Bond and Fox 2015). Figure 3 depicts a four-rating scale CCC after collapsing process took place. Note that the thresholds estimates value between categories have improved to more than 1.4 logits indicating a well-functioning scale.

E. Wright Map

This is the heart of Rasch Analysis and will be the premise of the instrument construct validity acceptance. It shows the logical hierarchy of item difficulty based on the conceptual theory put under test. A good item construct is evident when it is represented in the vertical direction. Being in horizontal direction shows redundancy of items measuring the same thing and is not desirable. Only when the item difficulty hierarchy is in place, then it is said the instrument has construct validity. Figure 4 depicts the item hierarchy map for each of the constructs for this study.

From the Wright Map it can be clearly seen that overall, items has a good hierarchical order with item measuring range of a mere 2.58, 2.87 and 2.55 logits for each of the construct respectively. It is also evident that a large number of items can be found along the continuum on which the majority of respondents' abilities fall. However, there are gaps in the hierarchical order for items belonging to EMEx construct which would require more items in between item rag_4 and rag_3 so that better meaning and measures can be achieved.

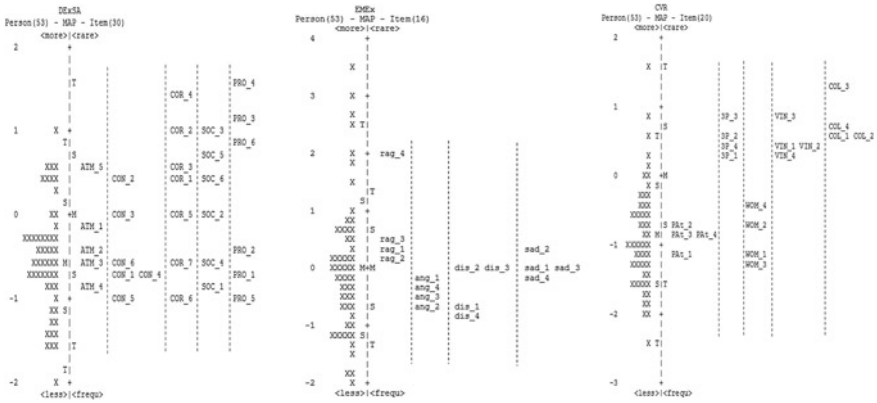


Fig. 4 Wright map for DExSA, EMEx and CVR

Other than that, there are also items that appear in horizontal order. Items are CON_1 & CON_4 from DExSA, dis_2 & dis_3, sad_1 & sad_3 from EMEx, and PAt_3 & PAt_4, VIN_1 & VIN_2, COL_1 & COL_2 from CVR. These items should be checked against the item dependency table (Table 4). If the pairs are not listed in the table, then the items is not a candidate for deletion.

Conclusion

In light of the aforementioned evidences, it seems that the application of Rasch analysis in refining research instrument facilitates the development of a more powerful tool for measurement. Rasch exposed the items to series of rigorous tests, producing measures with interval level data, which is an important requirement for high level analysis. As a result, the instrument yielded measures that have better fit and quality. Therefore, are more likely to produce more reliable and valid findings. Indeed, cross checks on the analyses confirmed that none of the suspected problematic items should be eliminated from the final instrument. However, analysis on category functioning curve suggests that the 6-point Likert rating scale should be collapsed to 4 categories to produce better measures. Therefore, the 66-items instrument with 4-point Likert rating will be used for actual study.

Other than that, the findings of this study would be very significant for organization in measuring customers' complaining behaviour as it provides a basis for a valid instrument construct that gives a better and true linear measure. This paper will also facilitate in adding new knowledge to existing literature in relation to consumer behavioral study.

Regarding future research, the logit measures obtained from this study should be imputed to other software such as smartPLS to investigate the relationship among the constructs as by doing so could help to produce better analysis and more

accurate results with interval level data. It cannot be denied that without the application of Rasch, good measurement is hampered in the absence of reliable instrument.

Acknowledgments The authors are grateful to all involved at the Faculty of Business and Management, Universiti Teknologi MARA, and to the Ministry of Higher Education Malaysia (MoHE) for the support given in carrying out this study. This study is funded by the FRGS Grant (Ref: FRGS/1/2015/SS01/UITM/01/1).

References

Abdullah, N., & Lim, B. K. (2013). Parallel circuit conceptual understanding test (PCCUT). *Procedia—Social and Behavioral Sciences*, *90*(InCULT 2012), 431–440.

Adams, J. S. (1963). Toward an understanding of inequity. *Journal of Abnormal Psychology, 67*(5), 422–436.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd Edn). Routledge.

Brentari, E., & Golia, S. (2008). Measuring job satisfaction in the social services sector with the Rasch model. *Journal of Applied Measurement, 9*(1), 45–56.

Conrad, K. J., Conrad, K. M., Dennis, M. L., Riley, B. B., & Funk, R. (2011). Validation of the substance problem scale to the Rasch measurement model. *GAIN Methods Report, 1*, 2.

De Battisti, F., Nicolini, G., & Salini, S. (2005). *The Rasch model to measure service quality*.

Fisher, W. J. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transaction, 21*(1), 2007.

Funches, V., Markley, M., & Davis, L. (2009). Reprisal, retribution and requital: Investigating customer retaliation. *Journal of Business Research, 62*(2), 231–238.

Ganglmair, A., & Lawson, R. (2003). Measuring affective response to consumption using Rasch modeling. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior, 16*, 198–210.

Green, K., & Frantom, C. (2002). Survey development and validation with the Rasch model. In *International conference on questionnaire development, evaluation, and testing* (pp. 1–30).

Grodin, J., & Blais, J. -G. (2010). A Rasch analysis on collapsing categories in item's response scales of survey questionnaire: Maybe it's not one size fits all. In *Annual meeting of the American educational research association* (pp. 1–29).

Huefner, J. C., & Hunt, H. K. (2000). Consumer retaliation as a response to dissatisfaction. *Journal of Consumer Satisfaction Dissatisfaction and Complaining Behavior, 13*(1), 61–82.

Idowu, O., Eluwa, A. N., & Abang, B. K. (2011). Evaluation of mathematics achievement test: A comparison between classical test theory (CTT) and item response theory (IRT). *Journal of Educational and Social Research*, *1*(4), 99–106. ISSN 2240-0524.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*(12), 1217–1218.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103–122.

Nor Irvoni, M. I., & Mohd Saidfudin, M. (2012). Students' perception towards quality library service using Rasch measurement model. In *2012 international conference of innovation, management and technology research (ICIMTR2012)* (pp. 668–672).

Nor Irvoni, M. I., & Rosmimah, M. R. (2016). Customer voice retaliation (CVR) construct verification : A Rasch analysis approach. In *Procedia economics and finance* (Vol. 37, pp. 214–220).

Pritchard, R. (1969). Equity theory: A review and critique*1. *Organizational Behavior and Human Performance, 4*(2), 176–211.

Salzberger, T. (2000). An alternative way of establishing measurement in marketing research—its implications for scale development and validity. In *ANZMAC 2000 visionary marketing for the 21st century: facing the challenge* (pp. 1111–1117).

Sekaran, U. (2003). *Research method for business—A skill building approach*. (J. Marshall, Ed.) (4th ed.). John Wiley & Sons, Inc.

Singh, J. (1988). Consumer complaint intentions and behavior : Definitional and taxonomical issues. *Journal of Marketing*, 93–107.

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17*(4), 396–408.

Yau, O. H. M., Chow, R. P., Sin, L. Y., Tse, A. C., Luk, C., & Lee, J. S. (2007). Developing a scale for stakeholder orientation. *European Journal of Marketing, 41*(11/12), 1306–1327.

Yoon, H. J., Song, J. H., Donahue, W. E., & Woodley, K. (2010). Leadership competency inventory: A systematic process of developing and validating a leadership competency scale. *Journal of Leadership Studies, 4*(3), 39–50.

A Preliminary Validity Study of Scoring for Japanese Writing Test in China

Jin-Chun Huang, Kai-Mei Zhang and Quan Zhang

Background

It is known to all, in China, there are many large-scale and nationwide tests of English language with plenty of corresponding Rasch-based research for the item analysis, test scoring, and equating (Shi-chun et al. 1995; Zhang 2011, 2012a, b, 2013a, b, 2014, 2015a, b; Mok and Zhang 2014, 2015); and the research has been playing the key roles in the field of English language testing since 1985; however, while the research is deepening in almost every aspect in the field of English test, such as test equating, validity studies, inter-rater and intra-rater inconsistency in terms of scoring for writing, tests of Japanese language in China remains inactive[1] though there are enormous tests of Japanese as well like Japanese professional forty-eight exam (NSS-4, NSS-8), College Japanese Test (CJT), Japanese Entrance Examination for National Higher Education and etc.

[1]The present author used keyword search "Rasch model, item response theory, Japanese test" to retrieve China's largest periodical full-text database—CNKI journals but did not find the use of Rasch model applied to study Japanese language tests.

J.-C. Huang (✉)
School of Foreign Languages, Lanzhou University of Technology,
Lanzhou 730050, China
e-mail: jchun6811@126.com

K.-M. Zhang
School of Foreign Languages, Beijing Normal University, Zhuhai 519090,
Guangdong, China

K.-M. Zhang · Q. Zhang
City University of Macau, SAR, Macau, China

Q. Zhang
Institute of Language Testing, Jiaxing University, Jiaxing, China

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_19

According to the statistics, the total number of Chinese studying Japanese in 2012 alone amounts to 1,046,490 people[2]; however, there is almost no corresponding yet significant empirical research (Wei 2013) based on statistical data. Since 1978, China also published the relevant research literature in Japanese teaching and testing, showing that research methods are mainly involved in the examination of lessons learned and ideas set forth and other non-empirical research-based, mostly concentrated in the Japanese major, Japanese college entrance exam, graduate entrance examination, but lack of breakthrough in terms of theoretical innovation, lack of research on the subject exam in terms of writing and speaking and lack of both technical support and theory model. On the other hand, the Internet's big data application also in some way restricts the practice and development of the Japanese testing in China (Wei 2013). On the whole, in the author's opinion, the problems inherent in Japanese testing practice in China nowadays is expected to improve in three aspects as follows: reliability and validity plus software ad hoc developed for this purpose; empirical studies and learners. The participation in PROMS2015, Fukuoka, Japan sparks my interest as well as activates the motivation to work in this direction. It is based on this, the present paper attempts to share, with Chinese teachers of Japanese, the validity study regarding the analysis of inter-rater inconsistency in our college Japanese writing test in an effort to promote such research in Japanese teaching and testing in China.

Research Design

Every June, in China, Test for Japanese Major, Band Four (NSS-4) is administered nationwide. NSS-4 covers the basic knowledge and use of Japanese language for Chinese sophomore students of Japanese majors, of which the focus is on the writing competence specified by the national syllabus for Japanese teaching stipulated by Minister of Education, China, wherein are categorized the writing types into narrating, explicit or letter writing with topics into campus life or social life specified like cultural value, hot events, funny story, a page from my diary, etc. However, there is one thorny problem for Chinese teachers of Japanese is that, though NSS-4 writing section is clearly specified, however, no detailed rubrics for writing have been revealed. This makes both teachers and students sort of puzzled about the scores given.

Motivated by this, the authors attempt to apply predictive validity study (Hughes 1989) and (Bachman 1981, 1982, 1996) to justify the current scoring practice of Japanese writing in China. According to Hughes and Lyle F. Bachman, a predicative validity is based on test results obtained from two parallel tests yielded by

[2]Source from the latest survey report 2013 by Japan Foundation, "Japanese foreign educational institutions.

the same group of subjects. These two tests are usually administered within a certain period of time interval; one test, or Test A usually a well-established test is used as the basal test and the other, Test B is used to be measured and the scores are to be predicted. In such a validity study, the focus is not on the comparison of the scores but on the ordering or the position of the subjects in the parallel tests. In doing so, the authors are also aware that many other potential factors such as motivation, learning method may affect the results in one way or other.

Subject and Tasks

In the present study, totally 40 students (of whom 6 are boys and 34, girls) randomly selected out of 64 college students of Japanese majors from two classes, Grade II, participated in a Japanese writing test or Test B. As required by NSS-4, the subjects were given 60 min to write a topic of "What impressed you most this semester" within 350 words.

To ensure the quality of the test results, Test B in the present study was administered as the final examination so that all the participants attached great importance to the test task. To validate this, all the subjects also took the real NSS-4, taken as Test A (as a basal test) a month later. Their NSS-4 total scores are used for further correlation analysis.

Raters and Rubrics

Three Chinese teachers of Japanese and three teachers of Japanese native speaker teachers are invited to grade the papers. These six raters used the rubrics made by the authors with references to the criteria based on Japanese Language Proficiency Test (JLPT), Test of Japanese for International Communication (J.Test) in Japan and on syllabus and curriculum for Senior Students of Japanese in China and Kawakami (2005), Komuro et al. (2005)and Kikuchi (1987). They fall into 4 bands as follows:

Band 4 Excellent, illustrating specific points, well-organized idea, good narration, logic and coherence, variety of choice of words, no grammatical errors, good handwriting, and appropriate style;

Band 3 Good, indicating some points not fully expressed, ideas are organized idea, logic and coherent, correct use of words, without serious grammatical errors, regular handwriting, and consistent style;

Band 2 Pass, showing the ideas are insufficient, poor coherence, inappropriate choice of words and obvious grammatical errors throughout the paper, legible handwriting but inconsistent style. The writer could basically make himself/herself understood in the written form

Band 1 Failure, showing the write unable to get the message across in the written form;

Research Purpose

The purpose of the present study is of threefold: to explore the differences inherent in the inter-rater inconsistency between Chinese and Japanese teachers, the high stake yielded from the official scores obtained from large-scale Japanese test and to promote such research method within Japanese testing in China.

Results and Discussion

Table 1 below shows the scores of Test B given by six raters and the scores of Test A obtained from Examination Authority. Table 2 indicates the correlation co-efficiency parameters demonstrated by the six raters.

Where Column Raters A, B, C are the scores given by Japanese raters and Column Rater D, E, F are the scores given by Chinese ones. The number 5 indicates the highest score and 1 the lowest score. Column test A shows the scores given by the Examination Authority of NSS-4.

As shown in Table 2 above, there exists in general good co-relationship between the scores given by both Japanese and Chinese raters. The scores given by all the raters turn out to be correlated with the scores given by the Examination Authority, and scores given by all the three Chinese teachers are highly corelated. This confirms in some way that our research design works, and no high stake may have been yielded from the official scores obtained from the large-scale Japanese test like NSS-4.

However, as can be observed, differences do exist in terms of the inter-rater inconsistency between Chinese and Japanese teachers as well as between Japanese teachers themselves. This can be illustrated in Rater B who behaved a little bit differently from all the other raters. For example, in one case, the negative corelation coefficiency (−0.01) with his Japanese counterpart, turns out, and in two cases, very lower co-relationship were observed with Chinese counterparts, and also the lowest co-relationship with Test A.

As we mentioned before, in such a validity study, there might be other potential factors such as motivation, learning method and some unknown factors like bias may affect the results in one way or other, so further efforts are needed to go on with the studies.

Table 1 The scores of Test B given by six raters and the scores of Test A obtained from Examination Authority

No.	Rater A	Rater B	Rater C	Rater D	Rater E	Rater F	Test A
1	3	4	1	1	4	1	48
2	4	4	5	3	3	3	69
3	3	3	4	3	3	3	79
4	3	4	4	3	1	3	65
5	5	4	5	4	4	5	82
6	4	3	4	4	4	3	79
7	3	4	4	3	3	2	50
8	3	3	4	5	4	3	72
9	3	4	2	3	3	2	60
10	3	4	3	3	4	2	67
11	4	4	5	4	4	5	92
12	3	4	4	3	5	3	78
13	3	3	4	3	3	3	66
14	4	3	5	5	5	5	79
15	3	3	3	2	2	2	50
16	4	5	4	4	3	3	72
17	3	4	4	4	4	3	65
18	4	5	4	3	3	3	77
19	2	4	3	2	2	2	54
20	5	4	5	4	4	4	79
21	3	3	5	4	4	4	85
22	3	4	4	4	4	3	66
23	5	5	2	4	4	3	65
24	4	3	3	4	3	2	58
25	5	3	5	3	3	4	71
26	3	4	4	3	4	3	73
27	3	3	3	3	3	3	67
28	5	3	3	2	4	2	62
29	4	4	4	3	3	3	63
30	3	4	4	3	5	3	65
31	3	3	3	4	5	1	46
32	5	5	4	5	4	4	67
33	5	4	5	3	5	4	72
34	3	3	5	4	4	3	60
35	3	3	4	4	3	3	45
36	4	4	5	4	4	4	69
37	3	3	5	4	3	3	74
38	3	2	3	3	2	2	37
39	4	3	4	3	3	3	53
40	2	3	2	1	2	1	46

Table 2 The corelation coefficiency matrix of the six raters and with test A

	Rater A	Rater B	Rater C	Rater D	Rater E	Rater F	Test A
Rater A	1	0.31	0.34	0.33	0.32	0.56	0.4
Rater B	–	1	−0.01	0.34	0.18	0.19	0.3
Rater C	–	–	1	0.55	0.25	0.8	0.62
Rater D	–	–	–	1	0.42	0.61	0.45
Rater E	–	–	–	–	1	0.33	0.4
Rater F	–	–	–	–	–	1	0.77
Test A	–	–	–	–	–	–	1

Conclusion

To conclude, the validity study first applied to Japanese writing test in China is successfully conducted. Though a very preliminary study, the significance lies in that it is the first time for Chinese teachers of Japanese to conduct such a research for a well-established Japanese writing test in China, a significant step towards promoting the use of statistical method to improve the reliability of Japanese writing test in China. There are limitations inherent in such a study, for example, the sample size is small and the research method needs improving. Ever since the participation in PROMS2015, Fukuoka, Japan, the authors have been aware that Rasch model or rather the multi-faced Rasch model (Bond and Fox 2001) would be the most appropriate to be used to deal with the problems. This would be no doubt the new direction guiding the Japanese language testing in China.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bond, T. & Fox, C. (2001). Rasch Model.

Hughes, A. (1989). *Testing for language teachers*. UK: Cambridge University Press.

Kawakami, M. (2005). A study of evaluation items for a proposal of evaluation criteria of many applications for writing: A fact-finding inquiry on teachers. *ICU Studies in Japanese Language Education, 2005*(2), 23–33.

Kikuchi, Y. (1987). The study of grading method in the writing test. *Japanese Language Education, 63*, 87–104.

Komuro, T., Mitani, S., & Murakami, K. (2005). The study of reliability and grading criteria on the writing test of EJU. *Janan.lnguage and culture, 2004*(3), 55–69.

Mok, M. M. C. & Zhang, Q. (2014). Constructing variables. Book of abstracts. *Journal of Applied Measurement* (Vol. 2). ISBN 978-1-934116-10-4. JAM Press P.O. Box 1283, Maple Grove, MN55311, USA.

Mok, M. M. C., & Zhang, Q. (2015). Constructing variables. Book of abstracts. *Journal of Applied Measurement* (Vol. 1). ISBN 978-1-934116-10-4. JAM Press P.O. Box 1283, Maple Grove, MN55311, USA.

Shi-chun, G., Li,W., & Zhang, Q. (1995). The Application of IRT to MET Equating. Guang Ming Press, Beijing.

Wei, Z. (2013). The review and status of domestic studies on Japanese language testing. *Foreign Language Research in Northeast Asia, 2013*(2), 77–83.

Ying, H. W. (2011). *A comparative analysis for the behaviors of scoring English writing test, 2011* (4), 27–32.

Zhang, Q. (2011). Towards better interaction between testing and teaching. *Keynote speaker at the 5th National TEFL/1st Mongolia TESOL Conference*, Ulaanbaatar, Mongolia, Oct 7–9, 2011.

Zhang, Q. (2012a). Towards international practice of language testing in China. Keynote speaker at the PROMS2012, Jiaxing, China, August 6–9, 2012.

Zhang, Q. (2012b). A pilot study based on Rasch into the appropriateness of the TOEIC bridge test for Chinese students: Status quo and prospect at the PROMS2012, Jiaxing, China, August 6–9, 2012.

Zhang, Q. (2013a). Invited keynote speaker at pacific-rim objective measurement symposium hosted by National Sun Yat-sen University, Kaohsiung, Taiwan, July 31–August 3, 2013.

Zhang, Q. (2013b). Rasch model and MET equating. Invited keynote speaker at assessment conference jointly hosted by the HONGKONG Institute of Education and Education Bureau, the Government of the KONG KONG Special Administrative Region, January 15–16, 2013.

Zhang, Q. (2014). Invited keynote speaker at pacific-rim objective measurement symposium hosted by Jiaxing University, Guangzhou, China, August 2–6, 2014.

Zhang, Q. (2015a). An overview of Rasch model: Status quo and prospect in China Invited Speaker at the 13th Asia TEFL International Conference hosted by Nanjing University: Creating the Future for ELT in Asia: Opportunities and Directions, Nanjing, China, November 6–8, 2015.

Zhang, Q. (2015b). The feedback function of learning assessment on teaching in the case of English subject, Invited Speaker at Conferência sobre " Avaliação diverificade-Impulsionar o sucesso n aprendizagem dos alunos" hosted by Direcção dos Serviços de Educação e Juvenude, Governo da Região Administrativa Especial de Macau, SAR, November 6–7, 2015.

Zhang, Q., et al. (2015a). Rasch model: Status quo and prospect in China. *Pacific-rim objective measurement symposium (PROMS) 2014 conference proceeding*. ISBN 978-3-662-47489-1, on Springer.com.

Zhang, Q., et al. (2015b). A Rasch-based approach for comparison of English listening comprehension between CET and GEPT. *Pacific-rim objective measurement symposium (PROMS) 2014 conference proceeding*. ISBN 978-3-662-47489-1, on Springer.com.

Verifying Measure of Supervisor-Rated Leader-Member Exchange (LMX) Relationship Using Rasch Model

Shereen Noranee, Rozilah Abdul Aziz, Norfadzilah Abdul Razak and Mohd Amli Abdullah

Background

An important and unique feature of leader-member exchange (LMX) theory is its emphasis on dyadic relationships. Yet, research on supervisor–subordinate relationships has shown convincingly that leaders do not behave consistently and similarly toward all subordinates. Instead, leaders form different quality relationships with their subordinates. High-quality LMX dyads exhibit a high degree of exchange in superior–subordinate relationships. Subordinates in these dyads are often given more information by the superior and reported greater job latitude. Lower-quality LMX relationships are characterized by more traditional "supervisor" relationships based on hierarchical differentiation and the formal rules of the employment contract.

Problem Statement

The leader-member exchange (LMX) theory of leadership focuses on the quality of relationships built between leaders and subordinates; LMX measures are designed to assess the quality of these relationships. Since the leader and subordinate are

S. Noranee (✉) · R.A. Aziz · N.A. Razak · M.A. Abdullah
Faculty of Business and Management, Universiti Teknologi MARA, Shah Alam, Malaysia
e-mail: shereen@puncakalam.uitm.edu.my

R.A. Aziz
e-mail: rozilah@puncakalam.uitm.edu.my

N.A. Razak
e-mail: norfadzilah0438@salam.uitm.edu.my

M.A. Abdullah
e-mail: amli_baharum@pahang.uitm.edu.my

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_20

jointly embedded in the relationship, it is reasonable to assume that their ratings of the relationship will converge to some reasonable extent. However, supervisor–subordinate relationships has shown convincingly that leaders do not behave consistently and similarly toward all subordinates. Instead, leaders form different quality relationships with their subordinates.

A good and valid instrument helps to determine how clearly leaders behave toward their subordinates. Therefore, the objective of this paper is to verify this instrument, the supervisor-rated LMX of their subordinates.

Literature Review

This section reviews the measurement of leader-member exchange (LMX) relationship. The purpose is to provide research background of the scale.

Leader-Member Exchange Relationship

Rooted in social exchange theory (Blau 1964) and the norm of reciprocity (Gouldner 1960), LMX focuses on the quality of the dyadic, interpersonal relationship between the supervisor and subordinate (e.g., Gerstner and Day 1997; Graen and Uhl-Bien 1995; Liden et al. 1997). Supervisors have been shown to give favourable treatment upon subordinates with whom they have high-quality LMX relationships. In return, subordinates have been shown to reciprocate favourable treatment upon their supervisors by engaging in extra role and extra task effort (e.g., Organ and Ryan 1995; Settoon et al. 1996).

Definitions of Leader-Member Exchange Relationship

Interpersonal interactions that employees secure in organizations have important implications on individual well-being, morale, effectiveness and competence, as well as organizational success and productivity (e.g., Shao et al. 2011; Podsakoff et al. 2009). The relationship allows employees to have more opportunity to perform citizenship behaviour for those people who are close to them, regardless of race, gender, or age (Bowler and Brass 2006). Interpersonal exchange relationships with supervisors are important that it ultimately determines how employees define and play their roles within the organizational context (Dienesch and Liden 1986). At the same time, subordinates are not passive, but rather proactive participants who would try their best to change their work environment. As high-quality exchange relationships develop, mutual internal goals and attitude similarities

between managers and employees are linked positively to job-related outcomes (Lo et al. 2006).

An important and unique feature of LMX theory is its emphasis on dyadic relationships. Yet, research on supervisor–subordinate relationships has shown convincingly that leaders do not behave consistently towards all subordinates (Dansereau et al. 1975; Graen 1976). Instead, leaders often form different quality relationships with their subordinates (Liden and Graen 1980).

High and Low Leader-Member Exchange Relationship

High-quality LMX dyads exhibit a high degree of exchange in superior–subordinate relationships and are characterized by mutual liking, trust, respect, and reciprocal influence (Dienesch and Liden 1986). Subordinates in these dyads are often given more information by the superior and reported greater job latitude. Lower-quality LMX relationships are characterized by a more traditional "supervisor" relationship based on hierarchical differentiation and the formal rules of the employment contract (Graen and Scandura 1987; Scandura and Graen 1984).

Subordinates in the in-group claimed to have more power as they receive more information, are more influential and confident, and obtain personal attention from their leaders as compared to the out-group subordinates (Liden et al. 2000). In-group members are willing to do extra tasks to which their leaders will reciprocate (Graen and Scandura 1987), but out-group members receive lesser attention and support from their leaders and thus, might see their supervisors as treating them unfairly. Literature (Dansereau et al. 1975; Brower et al. 2000) in the past has revealed that supervisors do differentiate between their subordinates in terms of the exchange. It is an advantage to be in a high-LMX as it is associated with higher trust, greater warmth, and support, and there is more frequent interaction between the members of the dyad (Dansereau et al. 1975; Brower et al. 2000). However, it should be noted that members having initially low LMXs are not necessarily poor performers in the work units. The initially low-LMX group clearly has the potential to consistently produce at higher levels, but it appears that they do not perceive higher performance as being worth the effort. After the one-on-one leadership intervention, the initially low-LMX group responds more positively to the new opportunities than do their colleagues (Scandura and Graen 1984).

It is important for the employees to know where they stand with their supervisors in an organization. To do this, employees need to compare themselves with other colleagues and assess how supervisors and colleagues react to them (Van Breukelen et al. 2006; Lamertz 2002). Employees are aware that with high-LMX, leaders are sensitive to subordinates' professional credentials and potential work contribution (Henderson et al. 2009), thus would lead to a greater increase in organizational commitment (Leow and Khong 2009).

Relational facets based on power are likely to be more important than instrumental facets. When supervisors demonstrate power capacities, such as treating

subordinates with respect, subordinates develop feelings of positive self-worth. Subordinates feel that the supervisors are treating them fairly while using the power, hence yielding positive effects on subordinates' organizational outcomes (Asgari et al. 2008).

Existing LMX research departs from previous ones in that, LMX has recently been related to behaviours such as organizational citizenship behaviours (Deluga 1998; Hui et al. 1999), task performance (Howell and Hall-Merenda 1999), turnover intention (Ansari et al. 2000), organizational outcomes (Omar 2001), and influence tactics (Liew 2003). LMX literature has found that subordinates in high-quality exchange relationships received more desirable assignments, more rewards, and had greater support from their supervisors. This is congruent with social exchange theory, where individuals who are engaged in high-quality relationships will behave in such a way that their exchange partner will also receive benefits (Lo et al. 2006).

When employees have high exchange relationships with their supervisors, they feel that this would lead to positive treatment by their supervisors. This would induce an obligation on the part of the followers to reciprocate positive treatment from leaders with extra role behaviours. Hence, these employees are motivated to help their leaders and, equally, the organizations achieve their goals (Chan and Mak 2012).

Reciprocity

The basic premise of LMX theory is that managerial actions that demonstrate positive regard for an employee create a desire on the part of the employee to reciprocate through behaviours that are highly valued by the manager (Settoon et al. 1996). Dienesch and Liden (1986) argued that both the theoretical model of LMX and the measure of this concept should be multi-dimensional and proposed a three factor model that included contribution, loyalty and affect.

As mentioned by Murry et al. (2001), the positive exchanges (Bernerth et al. 2007) are typically reciprocated with positive outcomes from the subordinates. Each member of the dyad has the other's best interest at heart and this is reflected in more supportive behaviour (Lo et al. 2006). Expectations and intention exist in helping behaviours. When a person does a favour by completing a colleague's task, there is indirectly an expectation of a return (Kandan and Ali 2010) and also could affect the rate of the return as well (Banki 2010). According to Blau (1964; cited in Bernerth et al. 2007), employees will try to balance between inputs and outputs of any of social transaction to stay out of debt and yield towards reciprocity. Even a leader expresses altruism for returning back the effort of his followers' effort.

Out of reciprocity of LMX, employees tend to repay what they feel they owe. A reason why a person stays committed to the organization might come from the feeling of reciprocity (Felfe et al. 2008). To balance the interpersonal interaction to

achieve equity, employees who want to repay a previous kindness may choose to do acts of citizenship behaviour (Poile 2010).

To sum up, LMX is a relationship-based approach to leadership that focuses on the two-way (dyadic) relationship between leaders and followers. It suggests that leaders develop an exchange with each of their subordinates, and that the quality of these LMX relationships influences subordinates' responsibility, decisions, and access to resources and performance.

Methodology

This research is based on a descriptive and exploratory study that employs the quantitative approach to obtain respondents' perspectives on LMX.

The methodology employed in this study is a cross-sectional correlational research, a quantitative and deductive research that describes the linear relationships between two or more variables (Sekaran and Bougie 2013). This study involved the gathering of data over a period of three months. The extent of researcher interference was minimal as the study was conducted in the natural environment of the organizations with the normal flow of work and non-contrived study setting. The data was analyzed using the statistical package in the Social Sciences Software (SPSS) version 20 and Rasch Model Measurement (WINSTEPS 3.72.3).

The population refers to the entire group of people, events, or things of interest that is to be investigated (Sekaran and Bougie 2013). Population is defined as a group of potential participants on whom the researcher wants to generalize the results of the study (Salkind 2014). The population of reference for this study included the total number of administrative officers, as supervisors, obtained from 20 public higher education institutions (referred to as public universities) working in various departments and units.

The sampling technique used was by stratified random sampling on 210 supervisors at public universities. A deliberate effort was made to obtain a representative sample by including administrative officers and their supervisors from 20 public universities in Malaysia. According to Sekaran and Bougie (2013), a sample of 100–500 is large enough to generalize the population. Furthermore, with reference to Krejcie and Morgan (1970), for a population size of 1919 (the round up number in the list is 1900), the sampling size should be 320. Thus, consistent with the above suggestions, the minimum sample size targeted in this study was set at 320.

On average, the response rate of survey questionnaires collected is approximately 30–50 % (Wallace and Mellor 1988). Using a response rate of 50 %, the number of questionnaires that need to be distributed should be a least 960 to achieve an effective sample of 320. Out of 979 pairs of questionnaires distributed,

only 577 (58.94 %) administrative officer respondent questionnaires' were returned. Meanwhile, for supervisor respondents, only 343 (35.04 %) were collected. However, there were only 210 (21.45 %) pairs that matched.

Stratified random sampling is the most efficient technique among all probability designs where all groups are adequately sampled and comparisons among groups are possible (Sekaran and Bougie 2013). Proportionate stratified random sampling technique was used whereby the population was first divided into meaningful segments (e.g., organizational categories); thereafter the number of subjects from each stratum would be altered, while keeping the sample size unchanged. The purpose of employing this type of sampling technique is to infer the data results that represent all local universities in Malaysia. Hence, proportionate sampling decision was made because it was suspected that more variability within a particular segment as well as administrative officers' designations (strata).

In this study, two levels of sampling techniques were conducted. First, the population was divided into organizational (i.e., university) categories. Stratified random sampling involves stratifying the elements along meaningful levels and taking proportionate samples from the strata. This technique is well-represented and more valuable and differentiated information would be obtained with respect to each group (Sekaran and Bougie 2013).

Second, for each important segment (i.e., subordinates and immediate supervisors) of the population, simple random sampling was used because every element in the population has a known and equal chance of being selected as a subject (Sekaran and Bougie 2013).

Findings of Study

An appraisal of data fit to supervisor-rated LMX relationship scale was conducted as a mean of verifying the measure. A total data point of 1770 evolved from 197 respondents on the 9 items analyzed. It produces a Chi-square value of 3309.74 with 1562 degree of freedom ($p = 0.000$, Table 1). This means that the overall fit to the measurement is good. For this set of data, it can be seen that the mean infit and outfit for item mean square are 1.01 and 1.00 respectively, very much as the expected value of 1.00 (Linacre 2011). Similarly, the mean infit and outfit person mean square are both at 1.02 and 1.00 respectively.

The mean Z standardized infit and outfit values are expected to be 0.0. As displayed in Table 1, the mean infit and outfit values for item's Z-standard are both 0.0, while the mean infit and outfit for the person's Z-standard are both −0.2. Since the values for the mean square and the Z-standard are very much close to the expected values, therefore, it can be said that the data for the actual research does fit the Rasch model reasonably well and appropriate analysis conducted can reveal the outcome of this research.

Table 1 Data fit to supervisor-rated LMX relationship scale

SUMMARY OF 9 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	740.2	196.7	.00	.11	1.01	.0	1.00	.0
S.D.	63.9	.5	.79	.01	.11	1.0	.11	1.0
MAX.	833.0	197.0	1.54	.13	1.19	1.7	1.16	1.5
MIN.	605.0	196.0	-1.29	.10	.87	-1.3	.83	-1.6

REAL RMSE	.12	TRUE SD	.78	SEPARATION	6.76	Item RELIABILITY .98
MODEL RMSE	.11	TRUE SD	.78	SEPARATION	6.94	Item RELIABILITY .98
S.E. OF Item MEAN = .28						

SUMMARY OF 197 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	33.8	9.0	1.53	.54	1.02	-.2	1.00	-.2
S.D.	4.7	.1	1.33	.09	.84	1.5	.82	1.4
MAX.	44.0	9.0	5.61	1.08	6.15	5.2	6.11	5.2
MIN.	18.0	8.0	-2.03	.44	.12	-2.8	.13	-2.8

REAL RMSE	.62	TRUE SD	1.18	SEPARATION	1.89	Person RELIABILITY .78
MODEL RMSE	.55	TRUE SD	1.22	SEPARATION	2.23	Person RELIABILITY .83
S.E. OF Person MEAN = .10						

INPUT: 420 Person 81 Item REPORTED: 197 Person 9 Item 5 CATS WINSTEPS 3.72.3
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .81
1770 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 3309.74 with 1562 d.f. p=.0000

Reliability of the Instrument

The process of reliability and validity is very much needed in evaluating the quality and reconstructing of a diagnostic tool. According to Jackson et al. (2002), reliability is the consistency of an instrument in measuring what is supposed to be measured. As for this study, it is desirable to verify the supervisor-rated LMX relationship measure to obtain a good and valid instrument that can help to determine how clearly leaders behave toward their subordinates. In terms of the item reliability, the final version of the measure produced 'excellent' item reliability (Fisher 2007) of 0.98 logit. It indicates that the probability of the difficulty levels of every item remaining exactly the same if the instrument were given to a different group of supervisors is high. Hence, the instrument holds an "excellent" position of not being dependent on the respondents.

The person reliability is identified as 'good' at 0.78 logit by Fisher (2007). In addition, the Cronbach Alpha (KR-20) Person Raw score test reliability is slightly higher at 0.81 logit. With the reliability at 0.81, if a similar set of instrument measuring the to obtain a good and valid instrument, then the likelihood of obtaining a similar pattern of ability in the person measure order table and the location of these supervisors on the person–item distribution map would be fairly similar (Aziz 2010).

Table 2 shows the comparison between cleaned and uncleaned values of supervisor-rated LMX relationship scale.

Table 2 Summary of comparison between cleaned and uncleaned values (9 measured item)

	Item	Person
Reliability	**0.98** (*0.97*) Excellent	**0.78** (*0.89*) Fair
MNSQ	**1.01** (*1.01*) Excellent	**1.02** (*1.01*) Excellent
Model error	**0.11** (*0.10*) Excellent	**0.54** (*0.36*) Very good
Separation	**6.76** (*5.56*) Excellent	**1.89** (*2.88*) Fair
Cronbach alpha	**0.81** (*0.91*)	
PCA variance measure	**48.3** (*42.7*) Fair	
Unexplained 1st contrast	**9.7** (*7.7*) Good	

Note a Cleaned values in bold, uncleaned values in italic, *b* rating scale instrument quality criteria (Fisher 2007)

In summary, the results showed a good reliability for both item and person measured at 0.98 (SE 0.11) and 0.78 (SE 0.54), respectively. The PCA of explained variance improved from 42.7 to 48.3 %, determining strong measurement dimension.

Discussion and Conclusion

The LMX measures were adopted from a 7-item (LMX7) construct of Scandura and Graen (1984) and additional 12 items were adopted from Bernerth et al. (2007). Several iterations were done by deleting the items identified as misfits. The better instrument was constructed, showing marked improvement across various fit statistics.

Quality control procedures have resulted in an item reduction from 19 to only 9 items, thereby producing a better instrument in measuring the supervisor rating of LMX of their subordinates. Referring to Fig. 1, after assessing the 9 items, the measurement can be divided into two categories; work-related and nonwork-related items. For work-related items, a supervisor expects a loyal employee to reciprocate through work contributions (Maden 2015). Whereas, for nonwork-related items, a

9-Measured Items
1. If he/she does something for me, I will eventually repay him/her. (**work-related**)
2. He/she can count on me to 'bail him/her out' at my expense when he/she really needs it. (**work-related**)
3. His/her relationship with me is composed of comparable exchanges of giving and taking. (**work-related**)
4. His/her opinion has an influence on me, and my opinion has an influence on him/her. (**work-related**)
5. Regardless of how much power I have built into my position, I would be personally inclined to use my power to help him/her solve problems at work. (**work-related**)
6. I understand his/her problems and needs well enough. (*non-work related*)
7. If I do something for him/her, he/she will return the favor at some point. (**work-related**)
8. He/she usually knows where he/she stands with me. (*non-work related*)
9. His/her working relationship with me is effective. (*non-work related*)

Fig. 1 9 measured items of supervisor-rated LMX relationship measurement

supervisor expects his/her employees to hold the feelings of respect and affect towards the supervisor (Long et al. 2015). This finding also could perhaps be attributed to the fact that Malaysian society is a hierarchical and relationship-oriented society (Ansari et al. 2004; Hwa et al. 2008).

Future Directions

This study developed a psychometrically sound LMX scale that captures the exchange process between leaders and members. A significant drop in the number of supervisor-rated LMX scale items which was originally from 19 to 9 items was an alarm. Even though, necessary steps were taken in ensuring that the items were of "good quality", the large number of item deletion needs to be investigated. Future investigation on this issue should be done to validate further on the 10 misfit items which could be related to high power distance, high collectivist, and high performance orientation of Malaysian culture.

However, the results using the revised instrument may yield awareness and understanding among subordinates how their supervisors perceive LMX of them; hence, necessary action can be executed by the employees, supervisors, and the organization, to improve LMX relationship among employees.

References

Ansari, M. A., Ahmad, Z. A., & Aafaqi, R. (2004). Organizational leadership in the Malaysian context. In D. Tjosvold & Leung (Eds.), *Leading in high growth Asia: Managing relationship for teamwork and change* (pp. 109–138). Singapore: World Scientific Publishing.

Ansari, M. A., Daisy, K. M. H., & Aafaqi, R. (2000). Fairness of human resource management practices, leader-member exchange, and intention to quit. *Journal of International Business and Entrepreneurship, 8*, 1–19.

Asgari, A., Silong, A. D., Ahmad, A., & Samah, B. A. (2008). The relationship between transformational leadership behaviours, organizational justice, leader-member exchange, perceived organizational support, trust in management and organizational citizenship behaviours. *European Journal of Scientific Research, 23*(2), 227–242.

Aziz, A. Z. (2010). *Rasch model fundamentals: Scale construct and measurement structure*. Kuala Lumpur: Perpustakaan Negara Malaysia.

Banki, S. (2010). Is a good deed constructive regardless of intent? Organization citizenship behavior, motive, and group outcomes. *Small Group Research, 41*(3), 354–375.

Bernerth, J., Armenakis, A., Feild, H., Giles, W., & Walker, H. (2007). Leader-member social exchange (LMSX): Development and validation of a scale. *Journal of Organizational Behaviour, 28*(8), 979–1003.

Blau, P. M. (1964). *Exchange and power in social life*. Transaction Publishers.

Bowler, W. M., & Brass, D. J. (2006). Relational correlates of interpersonal citizenship behaviour: A social network perspective. *Journal of Applied Psychology, 91*(1), 70–82.

Brower, H. H., Schoorman, F. D., & Tan, H. H. (2000). A model of relational leadership: The integration of trust and leader–member exchange. *The Leadership Quarterly, 11*(2), 227–250.

Chan, S. C., & Mak, W. M. (2012). Benevolent leadership and follower performance: The mediating role of leader–member exchange (LMX). *Asia Pacific Journal of Management, 29* (2), 285–301.

Dansereau, F., Graen, G., & Haga, W. (1975). A vertical dyad linkage approach to leadership within formal organizations: A longitudinal investigation of the role making process. *Organizational Behaviour and Human Performance, 13*(1), 46–78.

Deluga, R. (1998). Leader-member exchange quality and effectiveness ratings: The role of subordinate-supervisor conscientiousness similarity. *Group and Organization Management, 23* (2), 189.

Dienesch, R. M., & Liden, R. (1986). Leader-member exchange model of leadership: A critique and further development. *Academy of Management Review, 11*(3), 618–634.

Felfe, J., Schmook, R., Schyns, B., & Six, B. (2008). Does the form of employment make a difference? Commitment of traditional, temporary, and self-employed workers. *Journal of Vocational Behaviour, 72*(1), 81–94.

Fisher, W. P. J. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*(1), 1095.

Gerstner, C. R., & Day, D. V. (1997). Meta-Analytic review of leader–member exchange theory: Correlates and construct issues. *Journal of Applied Psychology, 82*(6), 827.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 161–178.

Graen, G. (1976). Role-making processes within complex organizations. *Handbook of Industrial and Organizational Psychology*, *1201*, 1245.

Graen, G. B., & Scandura, T. A. (1987). Toward a psychology of dyadic organizing. *Research in Organizational Behavior*.

Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *The Leadership Quarterly, 6*(2), 219–247.

Henderson, D. J., Liden, R. C., Glibkowski, B. C., & Chaudhry, A. (2009). LMX differentiation: A multilevel review and examination of its antecedents and outcomes. *The Leadership Quarterly, 20*(4), 517–534.

Howell, J. M., & Hall-Merenda, K. E. (1999). The ties that bind: The impact of leader-member exchange, transformational and transactional leadership, and distance on predicting follower performance. *Journal of Applied Psychology, 84*(5), 680.

Hui, C., Law, K. S., & Chen, Z. X. (1999). A structural equation model of the effects of negative affectivity, leader-member exchange, and perceived job mobility on in-role and extra-role performance: A Chinese case. *Organizational Behaviour and Human Decision Processes, 77* (1), 3–21.

Hwa, M. A. C., Muhamad, J., & Mahfooz, A. A. (2008). An investigation of the differential impact of supervisor and subordinate-rated leader-member exchange on career outcomes. *The Icfai University Journal of Business Strategy*, *V*(3).

Jackson, T. R., Draugalis, J. R., Slack, M. K., Zachry, W. M., & D'Agostino, J. (2002). Validation of authentic performance assessment: A process suited for Rasch modeling. *American Journal of Pharmaceutical Education, 66*(3), 233–242.

Kandan, P. A. L., & Ali, I. (2010). A correlation study of leader-member exchange and organizational citizenship behaviour: A public sector organization. *Journal of Global Business and Economics., 1*(1), 62–78.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educ Psychol Meas*.

Lamertz, K. (2002). The social construction of fairness: Social influence and sense making in organizations. *Journal of Organizational Behaviour, 23*(1), 19–37.

Leow, K. L., & Khong, K. W. (2009). The study of mentoring and leader-member exchange (LMX) on organisational commitment among auditors in Malaysia. *Sunway Academic Journal, 6*, 147–172.

Liden, R., & Graen, G. (1980). Generalizability of the vertical dyad linkage model of leadership. *Academy of Management Journal, 23*(3), 451–465.

Liden, R. C., Sparrowe, R. T., & Wayne, S. J. (1997). Leader-member exchange theory: The past and potential for the future. *Research in Personnel and Human Resources Management, 15*, 47–120.

Liden, R. C., Wayne, S. J., & Sparrowe, R. T. (2000). An examination of the mediating role of psychological empowerment on the relations between the job, interpersonal relationships, and work outcomes. *Journal of Applied Psychology, 85*(3), 407.

Liew, L. L. (2003). *Downward influence tactics: The impact of positive/negative affect, leader-member exchange, and gender*. Unpublished doctoral dissertation. Penang, University Science Malaysia, 27, 1–299.

Linacre, J. M. (2011). Rasch measures and unidimensionality. *Rasch Measurement Transactions, 24*(4), 1310.

Lo, M. C., Ramayah, T., & Hui, J. K. S. (2006). An investigation of leader-member exchange effects on organizational citizenship behavior in Malaysia. *Journal of Business and Management, 12*(1), 5.

Long, D. M., Baer, M. D., Colquitt, J. A., Outlaw, R., & Dhensa-Kahlon, R. K. (2015). What will the boss think? The impression management implications of supportive relationships with star and project peers. *Personnel Psychology, 68*(3), 463–498.

Maden, C. (2015). Linking high involvement human resource practices to employee proactivity: The role of work engagement and learning goal orientation. *Personnel Review, 44*(5), 720–738.

Murry, W. D., Sivasubramaniam, N., & Jacques, P. H. (2001). Supervisory support, social exchange relationships, and sexual harassment consequences: A test of competing models. *The Leadership Quarterly, 12*(1), 1–29.

Omar, F. (2001). *Downward influence tactics, leader-member exchange, and job attitudes*. Unpublished MBA thesis. Penang, University Science Malaysia.

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology, 48*(4), 775–802.

Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., & Blume, B. D. (2009). Individual- and organizational-level consequences of organizational citizenship behaviours: A meta-analysis. *Journal of Applied Psychology, 94*(1), 122.

Poile, C. (2010). *Asymmetric dependence and its effect on helping behaviour in work groups*. Doctoral dissertation, University of Waterloo.

Salkind, N. J. (2014). *Exploring research* (8th ed.). New Jersey: Pearson Education-Prentice Hall.

Scandura, T., & Graen, G. B. (1984). Mediating effects of initial leader-member exchange status on the effects of a leadership intervention. *Journal of Applied Psychology, 69*(3), 428–436.

Sekaran, U., & Bougie, R. (2013). *Research methods for business: A skill-building approach* (6th ed.). West Sussex, United Kingdom: Wiley.

Settoon, R. P., Bennett, N., & Liden, R. (1996). Social exchange in organizations: Perceived organizational support, leader-member exchange, and employee reciprocity. *Journal of Applied Psychology, 81*(3), 219–227.

Shao, P., Resick, C. J., & Hargis, M. B. (2011). Helping and harming others in the workplace: The roles of personal values and abusive supervision. *Human Relations*, 0018726711399940.

Van Breukelen, W., Schyns, B., & Le Blanc, P. (2006). Leader-member exchange theory and research: Accomplishments and future challenges. *Leadership, 2*(3), 295–316.

Wallace, R. O., & Mellor, C. J. (1988). Nonresponse bias in mail accounting surveys: A pedagogical note. *The British Accounting Review, 20*(2), 131–139.

Reliability and Validity Evidence of Instrument Measuring Competencies for Superior Work Performance

Nornazira Suhairom, Aede Hatib Musta'amal, Nor Fadila Mohd Amin and Adibah Abdul Latif

Introduction

Economic growth and current development in the field of hospitality and tourism industry have impacted culinary as one of the industry's important niche area. There are increased demands for high-skilled and competent culinary professionals working in the restaurant, catering, and hotels' sector. However, the high turnover rate of culinary professionals in the industry was one of the most confounding employment issues for the culinary profession. Additionally, recent studies have shown that the culinary graduates and young chefs are lack of knowledge, skills and abilities in performing their job. Thus, this study concerns on measuring the mastery level of competencies required for successful career in culinary profession. It is beneficial to have a comprehensive competency measurement instrument developed specifically for the profession can be used to provide data on competencies of current employees that we have in the industry. The establishment of a genuinely valid competency-based assessment approach can yield great benefit, not only to the professions, but to the whole community (Greenstein 2012; Gonczi et al. 1993).

Assessment should be precise, technically sound, producing accurate information for decision making in all circumstances (Dubois and Rothwell 2000; Stetz and Chmielewski 2015). The utilization of survey technique is communal among social science researchers as most of the data collection emphasizes on self-reported data. Nevertheless, the credibility of the instrument as a reliable and valid measurement tool is somehow flouted. It is important to consider the fact that identifying reliability and validity of an instrument is crucial for maintaining the accuracy of the instrument. In order to improve the survey instrumentation, Rasch measurement model was utilized to employ a data-driven model which is designed to measure

N. Suhairom (✉) · A.H. Musta'amal · N.F.M. Amin · A.A. Latif
Faculty of Education, Universiti Teknologi Malaysia, UTM Johor, Malaysia
e-mail: Nornadnazira@gmail.com

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_21

culinary professionals' self-assessment on their competencies. The Rasch measurement model was opted for this study because of its sophisticated approach to evaluate patterns of items responses and scale, and item performance (Linacre 2002; Bond and Fox 2007, 2015). Analysis using Rasch measurement model is a more sophisticated approach to evaluate patterns of items responses and scale, and item performance (Chen et al. 2014).

Methodology

The instrument testing was conducted among culinary professionals in 20 commercial kitchens of the 4-star and 5-star hotels in Peninsular Malaysia. The SC-SAT instrument was tested on 111 culinary professionals using a survey which had been conducted for three months.

Data Analysis and Findings

This section describes the data analysis and findings from the instrument testing. The questionnaire was analyzed using Winsteps software, software based on Rasch measurement model for reliability and validity test.

Demographic Profile

There are 111 respondents from 20 hotels in Peninsular Malaysia involved in the study. Table 1 shows the demographic profile of the respondents. About 61 respondents work with the 5-star hotels and 50 respondents work with the 4-star hotels. Majority of the respondents are Malay (83.8 %), followed by Chinese (9 %), Indian (4.5 %), and others (2.7 %).

According to gender classification, there are 74 (66.7 %) male and 37 (33.3 %) female culinary professionals. Among these 111 respondents, 52.3 % of them holds managerial level position (Chef de Partie post and above ranks) while 47.7 % works at nonmanagerial level (Commis and Cook). For each hotel's management, the title of job positions are varies; there are 16 types of job titles among respondents of the study. In terms of educational background, majority of them were Diploma holders (53.2 %), followed by high school graduates (43.2 %). There are three respondents who have Degree and one respondent who has a postgraduate educational accomplishment. For methods of culinary training and education attainment, 66 respondents (59.5 %) reported that they learnt culinary from culinary schools or institutions. Another 40.5 % of the respondents learnt culinary through

Table 1 Respondents' demographic profile

Demographic factors	Factors	*f*	%
Age (years old)	18–25	31	27.9
	26–35	57	51.4
	36–45	20	18.0
	<46	3	2.7
Job position	Executive Chef	3	2.9
	Executive Sous Chef	1	0.9
	Sous Chef	9	8.1
	Junior sous Chef	5	4.5
	Head Chef	4	3.6
	Pastry Chef	3	2.7
	Banquet Chef	1	0.9
	Demi Chef	10	9.0
	Chef de Partie	13	11.7
	Junior Chef de Partie	4	3.6
	Supervisor	1	0.9
	Chef	2	1.8
	Chef trainer	1	0.9
	Kitchen coordinator	1	0.9
	Cook	6	5.4
	Commis	47	42.3
Culinary experience	Below 5 years	29	26.1
	5–10 years	45	40.5
	11–15 years	19	17.1
	16–20 years	8	7.2
	21 years and above	10	9.0

experiences. About 83.8 % of the respondents have experience working in foreign countries. A large percentage of the respondents (84.7 %) do not have the MOSQ certification.

Person and Item Reliability and Separation Index

In the third instrument testing, the value for person reliability is 0.99 with person separation index of 8.78. Person reliability interpretation is equivalent with Alpha Cronbach (KR-20), which is 0.99. The person separation index value of 8.78 demonstrates that there are 9 levels of person ability that can be categorized in the instrument. With 111 person measuring 159 items in the SC-SAT instrument, Table 2 shows the value of item reliability is 0.94 with separation index of 4.02.

Table 2 Items reliability and separation index for each constructs in SC-SAT

Constructs	Total items	Item reliability	Separation index
Technical	64	0.95	4.15
Nontechnical	52	0.85	2.38
Personal quality	20	0.88	2.72
Physical state	3	0.73	1.66
Self-concept	6	0.70	1.54
Motives	14	0.96	4.80

The finding demonstrates that the probability of the SC-SAT instrument reliability when given to another group of sample with the same characteristics is 0.94. The separation index of 4.02 means that items in the SC-SAT can be categorize into four levels of difficulty. Table 3 shows the value of item reliability and separation obtained for the six constructs in SC-SAT. From the table, it can be seen that most of the constructs showed item reliability value that is greater than 0.70, ranging from 0.70 to 0.96. Physical state and self-concept shows the item separation index below 2 (1.66 and 1.54).

Based on Table 3, all of the constructs are accepted because the item separation indexes are equal to and higher than 2, which is considered as acceptable values except for physical state construct which need to be revised as the value of item separation is 1.72. However, the person reliability for physical state construct is 0.75 indicates a satisfying condition for further analysis.

Item Polarity

Based on Table 4, all of the correlation coefficient is positive for each of the constructs, showing the item ability to measure the competencies is valid (Linacre 2002). There are no items that need to be dropped based on polarity requirement because items are moving in one direction with the constructs.

Table 3 Person reliability and separation index for each constructs in SC-SAT

Constructs	Total items	Person reliability	Separation index
Technical	64	0.97	6.17
Nontechnical	52	0.96	5.14
Personal quality	20	0.89	2.89
Physical state	3	0.75	1.72
Self-concept	6	0.85	2.41
Motives	14	0.88	2.69

Table 4 Polarity of items

Constructs	PTMEA corr				Total items
	Min*	Item	Max*	Item	
Technical	0.47	OPS4	0.87	CRE2	64
Nontechnical	0.53	MGM5	0.74	MGM2	52
Personal quality	0.50	PS13	0.71	PS3	20
Physical state	0.85	PHY3	0.93	PHY2	3
Self-concept	0.79	SC1	0.87	SC3	6
Motives	0.55	MOT14	0.75	MOT9	14

Max* Maximum value; Min* Minimum value

Fit Statistics

Table 5 shows the summary of analysis of Item Fit and Person Fit for the instrument testing. Based on the table, ten respondents were identified to be the misfit person in measuring the six constructs of competency. They are person ID88, ID104, ID112, ID110, ID76, ID41, ID49, ID0, ID64, and ID40.

Accordingly, Table 6 shows the detailed analysis of Person Fit. The analysis shows that these people do not meet the requirement of Rasch model in analyzing the fit characteristics. Thus, suggested these people supposed to be removed from the analysis.

Further, Table 7 shows the item misfit for each of the items in the SC-SAT instrument. Nine items was found to have Infit MNSQ above 1.4 and ZSTD above 2. There is only one item with value of Infit MNSQ below 0.6 and ZSTD value below −2.00; which is SVC5 (I can apply stalls arrangement concept).

Figure 1 depicts the visual presentation of the bubble charts generated by the Rasch Analysis. Item that fits the models' expectations are located in the acceptable values between −2.0 and +2.0. Items which located on the right (>+2.0) are too erratic to be useful whereas items on the left (<−2.0) are too good to be true.

Table 5 Item and person fit for SC-SAT

Constructs	Item				Person			
	Infit MNSQ		Outfit MNSQ		Infit MNSQ		Outfit MNSQ	
	Min	Item	Max	Item	Min	ID	Max	ID
Technical	0.57	SVC5	1.35	OPS2	0.17	ID88	2.55	1D41
Nontechnical	0.64	CAR7	1.90	MGM5	0.06	ID104	3.11	ID49
Personal quality	0.67	PS1	2.02	PS13	0.09	ID104	3.09	ID49
Physical state	0.77	PHY2	0.89	PHY3	0.02	ID112	3.96	ID04
Self-concept	0.74	SC3	1.27	SC6	0.03	ID110	4.65	ID64
Motives	0.76	MOT5	1.48	MOT8	0.20	ID76	5.04	ID40

Table 6 Analysis of person misfit SC-SAT

Measure	Model SE	Infit		Outfit		PTMEA corr	Person
		MNSQ	ZSTD	MNSQ	ZSTD		
0.40	0.12	2.53	9.4	2.54	9.5	0.41	ID40
1.33	0.13	2.33	8.5	2.32	8.5	0.39	ID01
1.43	0.13	2.17	7.8	2.17	7.8	0.39	ID25
0.53	0.12	2.03	6.9	2.06	7.1	0.41	ID49
1.44	0.13	2.04	7.0	2.03	7.0	0.39	ID27
−0.24	0.11	1.90	6.3	1.87	6.1	0.43	ID65
0.50	0.12	1.87	6.0	1.80	5.6	0.41	ID57
2.17	0.13	1.81	6.0	1.80	6.0	0.38	ID41
4.21	0.17	1.80	5.8	1.73	4.8	0.31	ID68
2.03	0.13	1.70	5.3	1.70	5.3	0.38	ID114
0.21	0.12	1.66	4.8	1.61	4.5	0.42	ID38
1.88	0.13	1.62	4.7	1.64	4.8	0.38	ID48
Better fitting omitted							
1.52	0.13	0.39	−7.1	0.38	−7.2	0.39	ID73
2.00	0.13	0.28	−9.3	0.28	−9.3	0.38	ID60
0.00	0.11	0.27	−9.0	0.28	−8.9	0.42	ID30
1.93	0.13	0.24	−9.9	0.24	−9.9	0.38	ID99
1.80	0.13	0.20	−9.9	0.19	−9.9	0.38	ID104
1.91	0.13	0.20	−9.9	0.19	−9.9	0.38	ID75

Table 7 Analysis of item misfit for SC-SAT instrument

Measure	Model SE	Infit		Outfit		PTMEA corr	Item
		MNSQ	ZSTD	MNSQ	ZSTD		
0.34	0.15	1.78	4.6	1.85	4.9	0.60	PS13
0.82	0.15	1.69	4.1	1.78	4.6	0.61	MGM5
0.72	0.15	1.74	4.4	1.73	4.4	0.61	MOT8
−0.56	0.16	1.67	4.2	1.58	3.4	0.56	SC6
−0.64	0.16	1.51	3.3	1.62	3.5	0.56	PS2
0.78	0.15	1.54	3.4	1.59	3.7	0.61	ENT1
0.74	0.15	1.48	3.1	1.45	2.9	0.61	MOT7
−0.01	0.15	1.48	3.1	1.42	2.7	0.58	CAR15
0.29	0.15	1.47	3.0	1.41	2.7	0.59	PHY2
Better fitting omitted							
0.56	0.15	0.56	−3.7	0.57	−3.6	0.60	SVC5

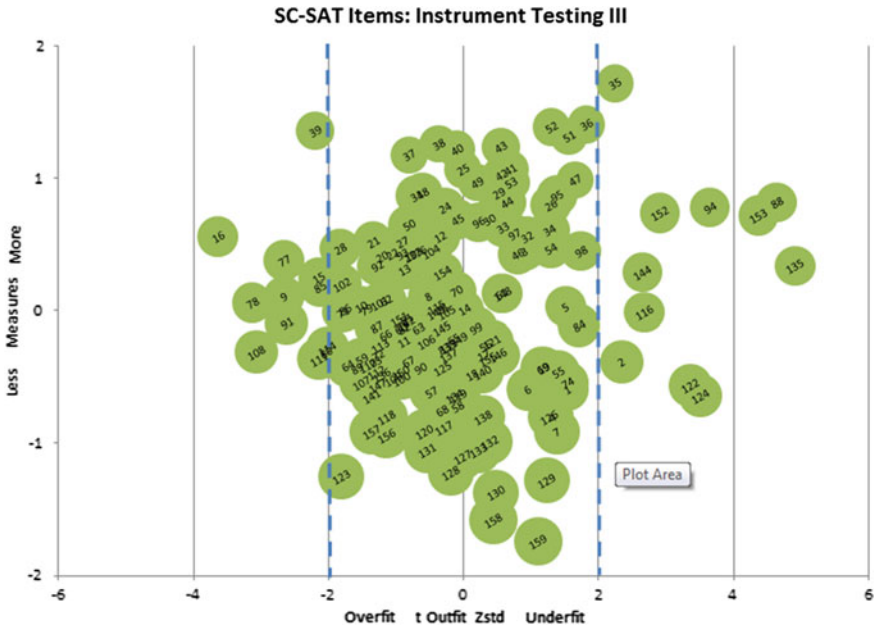


Fig. 1 Visual presentations of fit (quality control) for SC-SAT instrument

Item Dimensionality

Based on Table 8, the raw variance explained by measures is 41.8 %, whereas the unexplained variance in first contrast is 5.5 %.

Table 9 shows the value of raw variance explained by measures and the value of unexplained variance in first contrast for each constructs. The raw variance explained by measures ranging from 42.6 % (nontechnical) to 74.7 % (physical state). It is observed that this value is above the Rasch measurement requirement where the value must exceed 40 %. The range for the unexplained variance in first contrast is 7.1 % (nontechnical) to 13.3 % (physical state) also considered as an acceptable value below 15 %.

Table 8 Standardized residual variance (in eigenvalue)

Condition	Empirical (%)	Modeled (%)
Raw variance explained by measures	41.8	41.7
Unexplained variance in first contrast	5.5	14.3

Table 9 Standardized residual variance (in eigenvalue) for each constructs

Constructs	No. of items	Raw variance explained by measures (%)	Unexplained variance in 1st contrast (%)
Technical	64	47.1	8.3
Nontechnical	52	45.6	7.1
Personal quality	20	42.6	9.3
Physical state	3	74.7	13.3
Self-concept	6	64.0	10.6
Motives	14	47.9	11.7

Standardized Residual Correlation

The largest standardized residual correlation that is used to identify dependent items is displayed in Table 10. There are ten pairs of items that need to be revised because the value is more than 0.70, meaning that these items are highly correlated with each other.

Item Calibration

Analysis on item calibration was done to investigate whether appropriate rating scales are applied. The observed count is the number of times the category was selected across all items and persons. Based on Table 11, the scale used in the questionnaire is 5-point scale which described as 1: Not at All True of Me, 2: Slightly True of Me, 3: Moderately True of Me, 4: Very True of Me, 5: Completely True of Me. The scale number 4 "Very True of Me" is the most selected response from the respondents (48 %). The least response is for scale 1 (Not at All True of Me) with 0 % responses.

The observed average is normal and improved from negative to positive index. The index value starts from −0.38 to +3.26 logit. The category probability curve is shown in Fig. 2. Bond and Fox (2007) claimed that each of the rating categories should have a distinct peak in the probability curve graph. However, it can be seen that not the entire peak is clearly seen.

Further analysis on the calculation of Structure Calibration shows the values are not acceptable according to the requirement of $1.4 < SC < 5$ where $[-1.55 - (-2.42) = 0.87]$; thus collapsing is required between scale 1 and scale 2. After collapsing is done, the new Structure Calibration is improved. The observed average is increasing steadily and consistent as shown in Table 12.

The observed average, the average of logit positions modeled in the category is normal and enhanced from negative to positive index. The index value starts from

Table 10 Standardized residual correlations

Corr value	Item	Statement	Item	Statement
0.80	FIN4	I can write menu complete with financial management record	FIN5	I have the ability in implementing labor cost controls for my department
0.77	SCI3	I know the purpose of using food additives in food preparation	SCI4	I know the exact amount food additives to be used in cooking
0.73	OPS4	I know various cutting techniques in preparing foods	OPS6	I know how to apply appropriate cooking methods
0.72	HYG1	I have knowledge on the importance of hygiene for kitchen premises	HYG2	I comply to hygiene rules while in food handling and preparation
0.72	PHY1	I am fit for handling long hours events	PHY2	I am fit for handling heavy equipment in the kitchen
0.71	NUT2	I am skilled at developing menu based on special dietary needs	NUT3	I know the emergence of health products such as organic foods and genetically modified foods
0.70	OPS9	I have knowledge of producing products with original flavors	OPS 10	I am skilled at deploying size and portion of the food products
0.70	CAR13	I engage in activities that are directly linked to my performance	CAR14	I keep informed on affairs, structures, and processes in my profession
0.69	SVC3	I can identify type of buffet display and setting for catering services	SVC4	I can apply method of buffet display and setting dishes arrangement
0.68	INO3	I incorporate new ingredients to the recipe	INO4	I incorporate change in traditional cooking method

Table 11 Observed average at 5-point scale (12345)

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	OBSERVED SCORE	COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	74	0	-.38	-.57	1.11	1.17	NONE	(-3.75)	1
2	2	638	4	.09	.05	1.03	1.03	-2.42	-2.05	2
3	3	4587	26	.77	.83	.94	.94	-1.55	-.34	3
4	4	8476	48	1.87	1.82	.94	.94	.69	2.01	4
5	5	3873	22	3.26	3.30	1.06	1.05	3.28	(4.43)	5

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

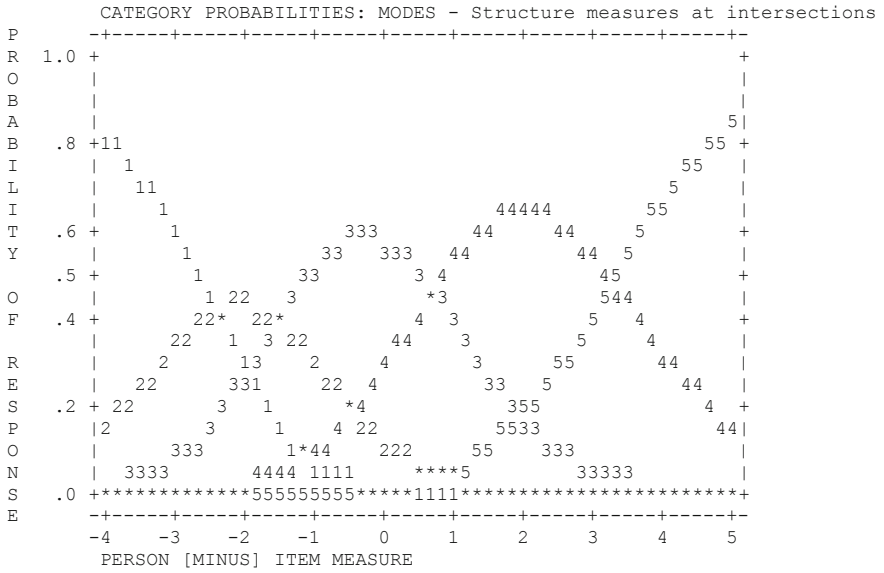


Fig. 2 Category probability curve at 5-point scale (12345)

Table 12 Observed average at 4-point scale!MediaObject ID="MO19">

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	712	4	-.80	-.86	1.04	1.04	NONE	(-3.49)	1
2	2	4587	26	-.08	-.02	.94	.94	-2.31	-1.25	3
3	3	8476	48	1.04	1.00	.94	.94	-.15	1.19	4
4	4	3873	22	2.45	2.48	1.06	1.05	2.46	(3.61)	5

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

−0.80 to +2.45 logit, demonstrating that it is increased by category value. The new category probability curve is shown in Fig. 3.

It can be observed from the figure that the entire peak can be seen distinctively. Further analysis on the calculation of Structure Calibration shows the values are well accepted according to the requirement of $1.4 < SC < 5$ where [−0.15 − (−2.31) = 2.16]. In an attempt to revise the categorization, the item and person reliability and separation index is reanalyzed for the calibrated scale of 11234. Table 13 shows the comparison of separation index value before and after scale calibration. Result shows that after category collapsing was done, the value of item separation index is increased from 4.02 to 4.03. On the other hand, the value of person separation index is maintained at 8.78. The value of mean person decreases from 1.81 to 0.98 (standard deviation 1.30).

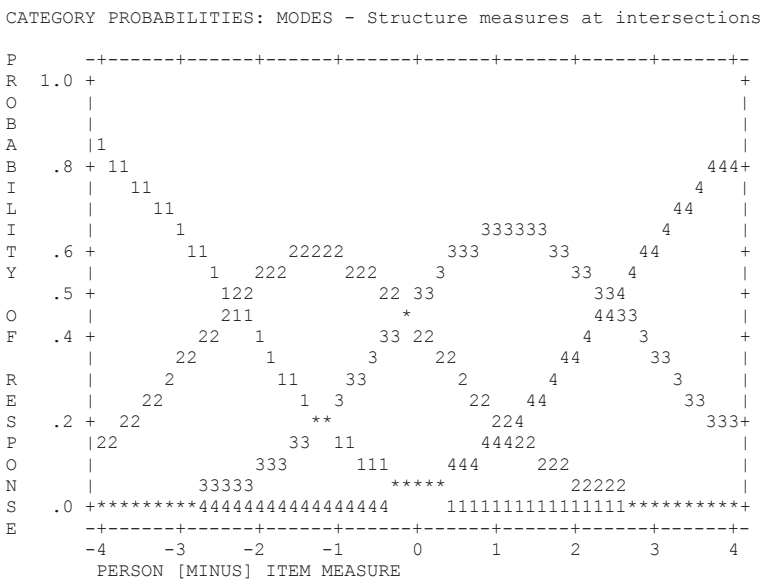


Fig. 3 Category probability curve at 4-point scale (1234)

Table 13 Comparison before and after scale calibration

Condition	Item		Person		Person mean
	Reliability	Separation	Reliability	Separation	
Before scale calibration	0.94	4.02	0.99	8.78	1.81
After scale calibration	0.94	4.03	0.99	8.78	0.98

The utilization of 5 likert-type rating scale is suggested for the next stage of instrument testing after taking into account that the changes in person and item separation appear to be meaninglessly small.

Differential Item Functioning (DIF)

It is crucial that the items in the designated SC-SAT instrument should not advantage (or disadvantage) culinary professionals from different groups. The differential item functioning (DIF) analysis is conducted to strengthen the psychometric evaluation of the instrument. The major purpose of DIF analysis is to identify whether there are biases exists among items in the SC-SAT instrument from the aspects of gender. To analyze DIF, Winstep perform the two-tailed t-test

Table 14 DIF analysis of SC-SAT items

Person class	DIF measure	Person class	DIF measure	DIF contrast	*t*	Item
1	−0.63	2	0.04	−0.67	−2.05	OPS2
1	−0.24	2	0.43	−0.67	−2.10	SVC3
1	1.95	2	1.27	0.68	2.34	SCI1
1	1.75	2	0.75	1.01	3.37	SCI2
1	1.40	2	0.75	0.65	2.17	SCI3
1	1.09	2	1.71	−0.62	−2.13	FIN4
1	0.70	2	0.04	0.66	2.07	QUA2
1	−0.27	2	0.87	−1.14	−3.63	EMO2
1	−0.63	2	0.04	−0.67	−2.05	EMO5
1	−0.84	2	−0.03	−0.82	−2.45	EMO6
1	0.42	2	1.10	−0.68	−2.23	ENT3
1	−0.31	2	−1.03	0.72	2.12	SC6
1	−0.31	2	−1.27	0.95	2.78	PS12
1	0.93	2	−0.88	1.81	5.52	PS13
1	−0.20	2	−1.03	0.84	2.47	PS14

1 Male, *2* Female

analysis to test the significant difference between the difficulty indexes. In DIF analysis, the cut off point is the critical t value within range of $+2.0 \geq t \geq -2.0$ and $+0.5 \geq$ DIF contrast ≥ -0.5 at 95 % confidence level.

Items which have DIF contrast value outside the range $\geq +0.5$ or ≤ -0.5 need to be revised after considering the *t* value. Results for the DIF Analysis of SC-SAT items based on gender are presented in Table 14. There are fifteen items which is detected as items that have bias between the two groups of male and female culinary professionals in the SC-SAT instrument. The analysis shows that most of the DIF measure for Person Class 1 (male) is smaller than DIF Measure for Person Class 2 (female), indicating that male culinary professionals more easy to endorse their self-reflections towards the competency items.

Item Targeting

The data were delved further to determine the Malaysian Culinary Professional's Competency Profile based on the SC-SAT instrument. Additional analysis were conducted to demonstrate the ability of the Rasch Analysis diagnoses in constructing the competency profiling based on the item difficulty and person ability. The heart of the Rasch Analysis is presented in Fig. 4, where the map of the person and items was displayed in tandem. The mean for all items are indicated as "M" (Item Mean) starts at 0.00 logit while the Person Mean (also marked as "M") is

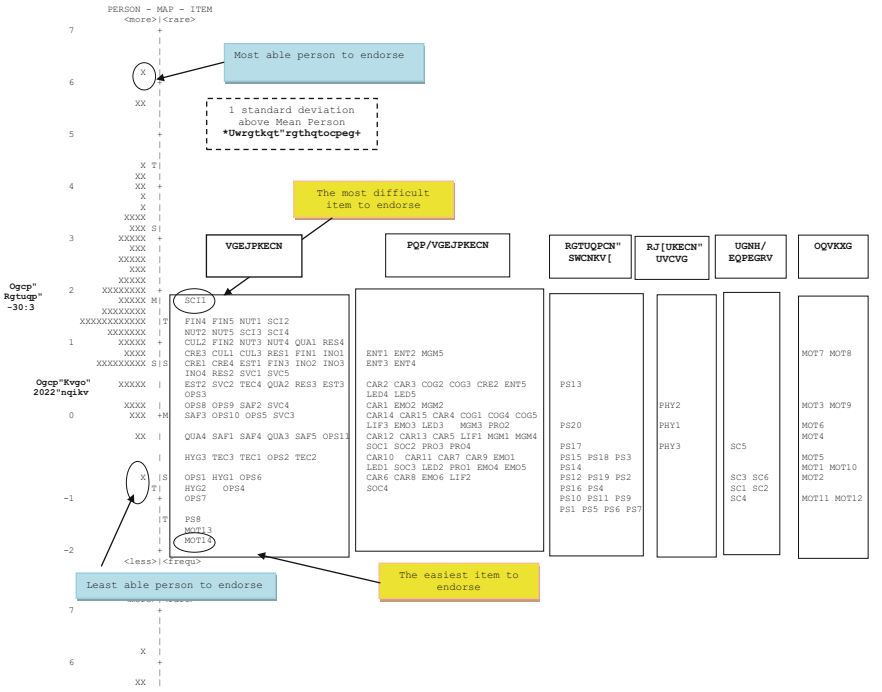


Fig. 4 Item-person wright map based on the SC-SAT

observed at +1.81. "S" is one standard deviation away from the mean, whereas "T" is two standard deviations away from the mean. From the Item-Person map, it shows that the Person Mean is above the Item Mean. Respondents' ability was arranged according to ascending order from the lowest to the highest ability in performing the items.

As shown in Table 15, the most difficult item is located at +1.72 logit and the easiest item is located at −1.74 logit with the standard deviation of 1.29, inferring the small spread within the data. Though the items still targeting at groups of person with moderate ability and below, there is an even distribution of persons according to their abilities along the logit scale. This shows that there is a slight improvement on item targeting. The most difficult item that respondents gave endorsement is item SCI1 from constructs technical competency "knowledge of cooking chemistry".

Table 15 Item difficulty level and person response level

	Item difficulty level		Person response level	
	Maximum	Minimum	Maximum	Minimum
Mean	0.00		+1.81	
Entry number	SCI1	MOT14	ID83	ID46
Logit value	+1.72	−1.74	+6.18	−0.62

Item M14 is the easiest items from motive construct "career as Chef brings the utmost satisfaction". There are 21 off target items with no respondents, which mean that these items are too easy. The maximum logit for person is +6.18 logit which is represented by person ID83, followed by two people at +5.67 logit (Person ID10 and ID82). The minimum logit for person is −0.62 logit (Person ID46).

From Fig. 4, there are two categories of items spanned along the positive and negative logit scale. There are 72 items above the mean (45 %) and another 87 items below the mean (55 %). This shows only 45 % of positive person response level, showing that these percentages of the respondents have perceptions to agree with the items. Accordingly, the findings demonstrate that they are capable in carrying out the competencies. Person ID83 who demonstrate the highest logit value (+6.18 logit) is a male, Chinese person who holds job position as a Sous Chef in a 5-Star hotel. He has been holding the position around 6–10 years. He is aged between 26 and 35 years old with more than 11–15 years of experience in the culinary industry. The person also has experience working in foreign country. Nevertheless, the person does not have MOSQ certification.

Discussion

As discussed earlier, Rasch measurement analysis was initially, conducted with 203 items, and conceptually ordered from low to high level of difficulty. It was concluded that a reliable linear, unidimensional scale of competencies for superior work performance was created using culinary professional views. With such detailed precision, these results mean that valid inferences could be made from the SC-SAT instrument. Since the scale data were shown to be reliable, valid inferences were drawn from the scale. Findings from the study have shed lights on the construct validity of the scales constructed. The study emphasized on six aspects of Rasch Analysis diagnoses which are (i) item and person reliability and separation index, (ii) item fit, (iii) item polarity, (iv) item dimensionality, (v) item calibration, and (vi) differential item functioning. The aspect of item targeting and consequently, competency profiling based on the SC-SAT instrument were discussed further. A closer look at the responses given by the culinary professionals in answering the SC-SAT may indicate which aspects specifically are sound and which may need attention to further developed their competence at work. The development of SC-SAT has put forward a better technique of competency measurement that is purposely developed for employees' professional development.

Assessment should include a spectrum of strategies where the process and products are emphasized. Assessment should be communicated, integrated in a day to day basis, stimulate thinking, build prior knowledge and construct meaning. Assessment results should be routinely revised and provide a proper database (Dubois and Rothwell 2000; Stetz and Chmielewski 2015). Formative assessment should be responsive in a way that it provides opportunities for self-reflections and revision. Hence, the betterment for model of assessment for professional

development is to provide feedback. Workers should receive feedback routinely, recognizing achievement beyond the scores.

Other than culinary profession, studies focusing on aspects of assessment of professional employees' competence and performance, addressing the question of self-assessment, and the means to assure more objective measurements of competence and performance were conducted widely among vocational profession (Winther and Klotz 2013), health professionals (Bashook 2005; Nicholson et al. 2012) information technology professionals (Azrilah et al. 2008), sales professionals (Lambert et al. 2014) and management professionals (Sisson and Adams 2013). The studies also attempts to develop applications of findings in identifying performance at workplace using a bevy of assessment methods.

Conclusion

Analyses of SC-SAT instrument items fully support its function as a useful measure of competencies. All items were analyzed and a minor modification has been made in order to achieve an adequate model fit. Data from the instrument testing provide evidence that the psychometric evaluations of the instrument are improved from one stage to another. The newly developed SC-SAT provides opportunity for culinary professionals to identify and measure their own competencies where the result can be used to identify how well they are doing. SC-SAT is considered as a norm-referenced measurement tool that is expected to possess a high degree of accuracy, discriminating those who perform excellently with those who are low performers, and functioning. Rasch Analysis has assisted the researcher in improving the quality of SC-SAT instrument, providing evidence to support the validity of the SC-SAT instrument. This study will be of better quality by implementing a number of improvements in a certain area such as in improving the item targeting.

References

Azrilah, A. A., Azlinah, M., Noor Habibah, A., Sohaimi, Z., Hamza, A. G., & Mohd Saidfudin, M. (2008). Development of Rasch-based descriptive scale in profiling information professionals' competency. *ISIT, 2008*, 1–8.

Bashook, P. G. (2005). Best practices for assessing competence and performance of the behavioral Health workforce. *Administration and Policy in Mental Health and Mental Health Services Research, 32*(5–6), 563–592. doi:10.1007/s10488-005-3265-z.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: LEA.

Bond, T., & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: LEA.

Chen, W.-H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item

characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research* 1–9.

Dubois, D. D., & Rothwell, W. J. (2000). *The competency toolkit* (Part II.). HRD Press.

Gonczi, A., Hager, P., & Athanasou, J. (1993). The development of competency-based assessment strategies for the professions. *National Office of Overseas Skills Recognition, Australia*, *Research P*.

Greenstein, A. (2012). *Assessing 21st century skills*. United States of America: Corwin.

Lambert, B., Plank, R. E., Reid, D. A., & Fleming, D. (2014). A competency model for entry level business-to-business services salespeople. *Services Marketing Quarterly, 35*, 84–103.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Nicholson, P., Griffin, P., Gillis, S., Wu, M., & Dunning, T. (2012). Measuring nursing competencies in the operating theatre: Instrument development and psychometric analysis using item response theory. *Nurse Education Today*, 6–11. doi:10.1016/j.nedt.2012.04.008

Sisson, L. G., & Adams, A. R. (2013). Essential hospitality management competencies: The importance of soft skills. *Journal of Hospitality & Tourism Education, 25*(3), 131–145. doi:10.1080/10963758.2013.826975.

Stetz, T. A., & Chmielewski, T. L. (2015). *Competency modeling documentation* (pp. 1–22).

Winther, E., & Klotz, V. K. (2013). Measurement of vocational competences: An analysis of the structure and reliability of current assessment practices in economic domains. *Empirical Research in Vocational Education and Training, 5*(1), 2. doi:10.1186/1877-6345-5-2.

Writing Assessment in University Entrance Examinations: The Case for "Indirect" Assessment

Kristy King Takagi

English language programs in Japan, and around the world, often use placement testing to gauge students' English proficiency levels and then place students into classes at those levels. English teachers and program administrators typically favor placement testing because it allows for more efficient teaching and because class placement affects and matters to students. However, placement testing for writing classes can be burdensome and time-consuming because a typical approach in language programs is to obtain writing samples from students and ask teachers to rate the samples. In light of this burden, and the problems associated with the rating of writing samples, the focus of this paper is to examine whether an objective multiple choice test of writing knowledge could serve as a supplement to or substitute for the typical rating of writing samples for English writing class placement.

As Hamp-Lyons (1991) pointed out in "Basic Concepts," the preferred method of testing writing has changed over time. Evaluating or rating of writing samples, often referred to as direct assessment, was typical procedure until about the 1940s, but ideas about writing assessment began to change. In the 1950s and 1960s, multiple choice tests of knowledge about writing, often referred to as indirect assessment, came into favor, thanks to the structuralist-psychometric ideas popular then (p. 7). But the 1970s saw a return to "language as communication," so that writing was once again assessed via rating of writing samples (p. 9). As one result, in 1986, the Test of Written English (TWE) was included on the Test of English as a Foreign Language (TOEFL), and was intended to be a direct measure of writing.

Why have objective, multiple choice writing tests not been regarded as appropriate tests of "language as communication"? Hamp-Lyons (1991) said that she did not believe that the skills needed on such tests "represent what proficient writers do" (p. 7). Similarly, Kroll (1998) said that "few in the teaching community feel

K.K. Takagi (✉)
The University of Fukui, Famille-Ai 3-201, 4418 Oyama-Cho, Machida, Tokyo 194-0212, Japan
e-mail: kjktakagi@hotmail.com

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_22

comfortable making credible claims about writers' skills on the basis of any sort of indirect measure on its own" (p. 221). Stansfield and Ross (1988) also discussed the discomfort expressed by some writing scholars over multiple choice "indirect" measures of writing. Reasons for not using objective tests of writing focus on what feels "right" in assessing writing ability, but statistical evidence for these beliefs and feelings is much harder to come by. Nevertheless, the result of discomfort with objective writing measures is that the usual method of testing writing since the 1980s, especially for placement purposes, has involved obtaining and evaluating a writing sample from students.

Rater assessment of writing samples, then, has generally been the preferred method of testing writing, but is it without problems? Not at all. There are a variety of problems associated with rating essays. First of all, the time involved can be considerable, a fact well known to writing teachers and administrators. Another difficulty is the choice of rubric or rating scale. While raters may work faster with holistic scales, the analytic rating would tend to be generally more reliable because it is comprised of a number of scores, instead of only one.

After English program administrators decide which type of scale to use, they still must choose from an assortment of rubrics, such as the TWE rubric, the Constructed Response Rubrics created by the makers of the Comprehensive English Language Test (CELT), Brown and Bailey's (1984) analytic scale, the ESL Composition Profile (Jacobs et al. 1981), and many other less known rubrics created by a wide variety of English language programs. Some of these have been evaluated and validated, but most have not. Clearly, there is a need to examine the rubrics; as Davidson (1991) pointed out, there can be serious problems in the calibration of rating scales. In examining a rubric used for rating essays for a high-stakes Japanese university entrance examination, for example, I found that the rubric favored for many years by the administrator was problematic (Takagi 2014). For example, all levels of rating categories were not used by the raters, and the threshold calibrations were too close together at three of five points. As Bond and Fox (2007) said, such small differences between steps indicate that each step was not clear in defining "a distinct position in the variable" (p. 224). Raters were not able to use the rubric completely or with precision.

Another potential problem with rating of writing samples is lack of rater agreement. If traditional methods are used to analyze results, strong interrater reliability is necessary; therefore, raters need training and practice (for a description of rater training for the TWE, see Stansfield and Ross 1988, p. 177). However, there tends to be little time available for such preparation; as a result, raters often are not experienced or trained, and, not surprisingly, their ratings of the same compositions frequently differ. In addition, raters can have individual problems in being too "safe" in using the rating scale, and therefore, overly predictable, or much worse for the measurement, unpredictable, and inconsistent.

When essays are being assessed for class placement, this problem of rater disagreement requires a procedure to resolve serious discrepancies. Some programs

follow a recommended procedure of asking a senior rater to make the final decision (Brown and Bailey 1984), but, because senior raters are not necessarily the best raters, this solution is not without problems (Takagi 2014). Other programs do not address rater disagreement at all, and simply average or total ratings. In conclusion, then, the often preferred "direct" method of rating compositions for writing assessment and placement can be fraught with difficulties, and therefore prone to inconsistencies and error.

Given the many difficulties associated with "direct" rating of compositions, surely objective writing tests can be useful tools for writing placement. Even Hamp-Lyons and Kroll admit their value. Hamp-Lyons (1991) said that these tests have correlated "fairly highly with measured writing ability" (p. 7), and Kroll (1998) said that these tests have been "valid predictors of writing ability as measured by their correlation with actual writing samples" (p. 221). Even vocal opponents to objective writing tests recognize the evidence for using them as tests of writing ability. In addition, the supposedly clear distinction between "direct" and "indirect" writing assessment is arbitrary; as McNamara (2006) said, testing is "a procedure for drawing inferences about the unobservable; it is necessarily indirect and uncertain" (p. 32). In other words, all testing is indirect in that we are attempting to measure an unobservable and latent variable, such as writing ability. Rather than creating such arbitrary distinctions between types of writing assessment, we should aim to create and validate the best writing tests possible, tests that include objective measures of writing knowledge.

In line with this aim of creating a useful objective measure of writing knowledge, I developed and pilot-tested the Sentence Form Test (SFT). There have been predecessors to this kind of writing test. Brown (1996) described the ESL placement test used at the time at the University of Hawaii as having two parts, a Writing Sample (composition), and a multiple choice proofreading test called The Academic Writing Test (p. 283). In addition, the Structure and Written Expression (Section 2) of the TOEFL, still included in the TOEFL PBT (paper-based test), is also believed to be a useful objective measure of writing skill; the Educational Testing Service (ETS) claims that it "measures the ability to recognize language appropriate for standard written English" (ETS 2016). According to Stansfield and Ross (1988), structure and written expression scores have correlated at about 0.70 with the TWE (p. 164); in other words, this objective measure of writing had a strong relationship to ratings of writing samples.

The SFT also could be called a proofreading test, but it is more precisely a test of sentence form, which tests ability to recognize correct versions of the four traditional types of sentences (simple, compound, complex, and compound-complex), the building blocks of all English writing. For each test item, students were asked to find the one incorrect sentence out of four choices. Incorrect sentences all had a serious structural error (often called major error), such as subject-verb agreement error, fragment, comma splice, etc. Such errors indicate an insufficient grasp of the language; therefore, the test was designed with the assumption that students who

recognize major errors on the SFT have a more complete knowledge of English sentences and of English writing than those who cannot do so. The test was purposely timed because, as Ellis and Barkhuizen (2005) explained, such tasks tap into implicit knowledge more than untimed tasks. In addition, the SFT was designed to tap into and account for learners' implicit knowledge, the main goal of SLA research, according to Ellis and Barkhuizen (2005). They noted that grammaticality judgment tests are very useful for "investigating specific grammatical structures that often prove difficult, or even impossible, to elicit in learner production" (p. 20). Some structures on the SFT are not usually produced by students who use English as a second or foreign language, so the SFT should work well in evaluating their knowledge of these structures.

In conclusion, then, the SFT was designed to test knowledge of English sentence structure, and it was hypothesized that this knowledge would be closely related to knowledge of English writing (as measured by performance on a writing task). As already noted, similar "indirect" tests did correlate well with ratings of writing samples; therefore, it is hypothesized that the SFT will also do so, and therefore tap into the same construct of writing ability that a composition task taps.

If the SFT and other tests like it can be shown to tap into the same construct (of writing ability) that a composition task does, then writing programs could have more options regarding writing placement test administration. For example, programs in which time and personnel abound could add another measure; multiple measures would make the writing assessment more reliable. On the other hand, if programs had no writing placement, little time, or a large number of students, then a test like the SFT alone could be used for writing placement purposes. Therefore, the specific purpose of this paper is to evaluate the SFT, especially in relation to the essay ratings given concurrently, in order to: (a) evaluate the SFT as a test and (b) determine to what extent the SFT taps into the same writing ability construct that a composition task does.

Research Questions

It is hypothesized that the SFT will work well as a writing placement test for a university EFL or ESL program, and that it will tap into the same construct of writing ability that the writing section of the test taps. In order to test this hypothesis, the following research questions were posed:

Research Question 1: Does the SFT work well as a writing test in that test items match student ability, create a useful spread of student ability, display acceptable fit values for the Rasch model, demonstrate acceptable reliability, and are unidimensional in measuring one construct?
Research Question 2: Can the SFT be shown to tap into same construct of writing ability that a composition task taps?

Method

Participants

Fifty freshman students at a women's college in Tokyo, Japan, took a writing placement test after finishing one year of composition study. Forty-five students were Japanese, and five were Chinese exchange students, all approximately 19 years of age. At the time the writing placement test was administered, students had all studied English for approximately seven years: six years in junior high and high school, and one year in college. Most Japanese students had not studied or lived overseas. Their language proficiency varied from basic to high intermediate; specifically, their Pre-TOEFL ITP scores from January of the same year ranged from a low of 293 to a high of 480 (mean of 383.20 and standard deviation of 36.40). Results of the writing placement test were to be used to place students into second-year writing classes.

Materials

The writing placement test included two sections. The first section was a composition in which students were asked to write about what they had learned in their first-year at university. Since instruction in the five levels of the first-year writing classes varied considerably, students were allowed to write either an essay or a long paragraph of about 250 words. The time limit for the writing assignment was 40 min. The second section of the test was the Sentence Form Test (SFT). On each test item, they were asked to find one incorrect option out of four example sentences. Students were given 20 min to complete the test. The following are directions and an example item included on the test:

Read the four sentences for each question. One of the sentences has an important mistake. Write the letter of the sentence with the mistake.

EXAMPLE

(a) I am late.
(b) I late.
(c) I was late.
(d) I was not late.

Answer: b

Scoring

The SFT answer sheets were quickly scored (in about 15 min). The 50 compositions were scored by four raters, all university teachers in Japan. Three of the raters

used Brown and Bailey's (1984) analytic scale, and two raters used a holistic scale used by the English program at Hawaii Pacific University. The analytic raters were all university EFL composition teachers; one was American, the second was British, and the third was Chilean. The holistic raters were Americans teaching college EFL. The process of scoring the essays took approximately four hours. The analytic scale yielded a maximum of 100 (with a maximum of 20 each for (a) organization; (b) logical development of ideas; (c) grammar; (d) punctuation, spelling, and mechanics; and (e) style and quality of expression). The holistic scale was originally based on a 0–10 point scale but was converted to a 100-point scale.

Procedures and Data Analysis

In order to answer the research questions, a number of statistical analyses were used. In answering the first question (determining whether the SFT generally worked well as a writing placement test), I conducted a Rasch analysis using Winsteps, version 3.90.0, in order to examine the variable map, and the fit and difficulty of SFT items. I examined test reliability with the Rasch model, as well as through traditional methods for assessing internal consistency. I also investigated unidimensionality of the SFT by examining a bubble chart pathway plot produced with Winsteps.

In answering the second research question (determining whether the SFT taps into the construct of writing ability tapped by writing task ratings) I first examined intercorrelations of SFT scores and writing task ratings because correlation coefficients indicate the degree to which measures "tap the same construct" (Stansfield and Ross 1988, p. 168). I then examined unidimensionality through principal components analysis.

Results

Research Question 1: Does the SFT work well as a writing test in that test items match student ability, create a useful spread of student ability, display acceptable fit and difficulty, demonstrate acceptable reliability, and are unidimensional in measuring one construct?

Descriptive statistics for essay ratings and test scores are shown in Table 1.

Figure 1 shows a variable map for the SFT produced by Rasch Analysis. The software used was Winsteps, Version 3.90.0, developed by J.M. Linacre. The 20 test items are on the right side, with most difficult (item 16) at the top, and least difficult at the bottom. The 50 students are on the left. The student with highest ability (student 22) is at the top, and least able students are shown at the bottom. The variable map shows that test items are mostly a good match for the students,

Table 1 Mean scores and statistics for essay ratings and SFT

Measure	HR1	HR2	AR1	AR2	AR3	SFT
M	48.40	59.96	68.52	67.34	55.94	9.44
SD	17.77	14.70	10.18	10.21	15.90	3.91
Skewness	0.50	0.14	−0.34	0.03	−0.21	0.41
SE of skewness	0.34	0.34	0.34	0.34	0.34	0.34
Kurtosis	−0.17	−0.29	0.16	−1.01	−0.52	−0.44
SE of kurtosis	0.66	0.66	0.66	0.66	0.66	0.66

Note $N = 50$ for all ratings. *HR1* holistic rater 1 scores; *HR2* holistic rater 2 scores. *AR1* analytic rater 1 scores; *AR2* analytic rater 2 scores; *AR3* analytic rater 3 scores. *SFT* Sentence form test scores. Possible score range for essay ratings is 0–100, and for SFT, 0–20 points

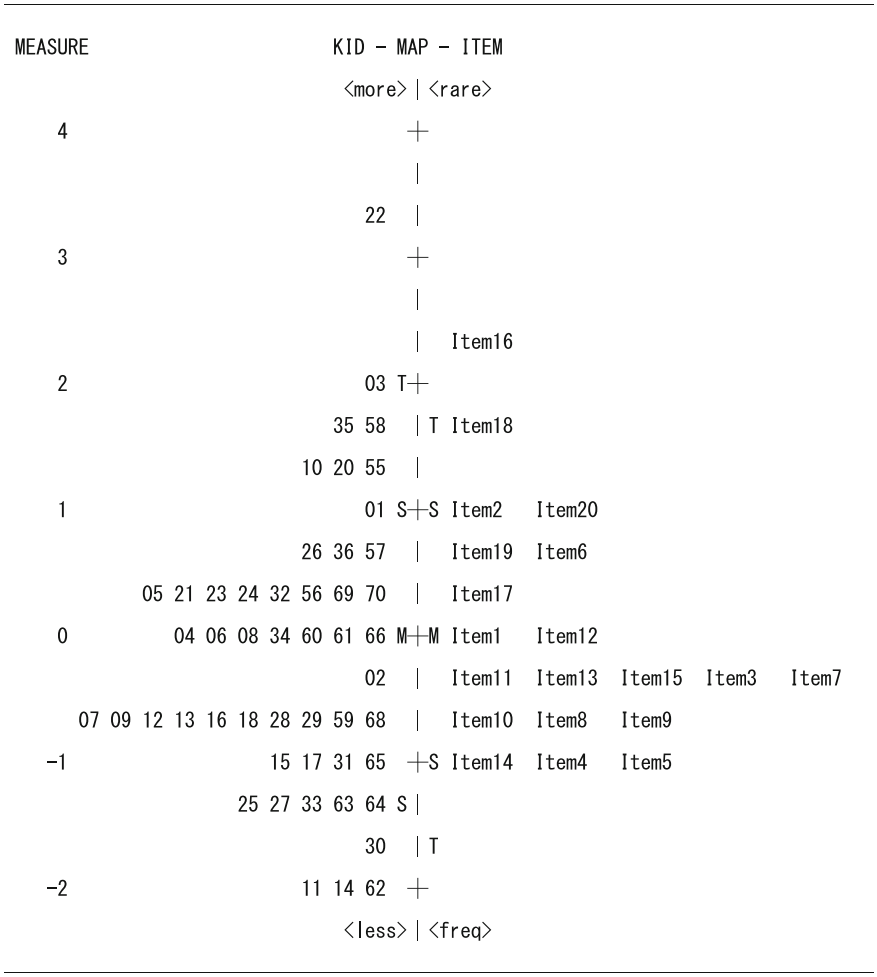


Fig. 1 Variable map for person ability and item difficulty of the SFT

and there is a good spread of item difficulty. Ability levels of students are spread out in a way that is useful for placement into writing classes.

The fit and difficulty of test items were assessed using the Rasch model. Table 2 shows the Infit Mean Square (Infit MNSQ), Outfit Mean Square (Outfit MNSQ), and Measure (indicating difficulty) for each item. According to Bond and Fox (2007), mean square infit and outfit values for a multiple choice high-stakes test should range from 0.8 to 1.2, and for a "run of the mill" multiple choice test, from 0.7 to 1.3 (p. 243). These "run of the mill" values would be acceptable for a writing class placement test.

Table 2 shows (in the Measure column) that the items generally move from easier to more difficult items, as was intended in the test design. However, this progression is not perfect, and some items do not follow the intended pattern. For example, items 2 and 6 are not as easy as later items, while item 14 is easier than intended. The Infit and Outfit MNSQ columns show that almost all test items fit the model well, though both infit and outfit values for Item 20 are too high.

Results from the Rasch Analysis revealed that the SFT item reliability was 0.85, and the student reliability was 0.73. Other traditional methods for assessing internal consistency were also used, for purposes of comparison. As recommended by Brown (1996, pp. 194–203), the split-half method adjusted by using the Spearman-Brown prophecy formula was employed. Since the test was designed to be progressively

Table 2 Rasch model descriptors of SFT test items

Items	Infit MNSQ	Outfit MNSQ	Measure
Item 1	0.89	0.83	−0.09
Item 2	0.86	0.75	0.86
Item 3	1.01	0.92	−0.19
Item 4	0.85	0.76	−1.09
Item 5	0.96	0.97	−0.99
Item 6	1.17	1.38	0.63
Item 7	0.84	0.79	−0.38
Item 8	1.04	0.99	−0.68
Item 9	0.94	0.93	−0.68
Item 10	0.73	0.64	−0.78
Item 11	1.07	1.07	−0.48
Item 12	0.94	0.96	−0.09
Item 13	1.09	1.07	−0.19
Item 14	0.84	0.73	−1.09
Item 15	1.02	0.93	−0.38
Item 16	0.87	0.65	2.24
Item 17	1.17	1.41	0.21
Item 18	1.28	1.30	1.67
Item 19	1.08	1.21	0.52
Item 20	1.34	1.61	0.97

Note $N = 50$

more difficult, splitting it into two halves by dividing odd-numbered from even-numbered items made sense. The resulting Spearman-Brown Coefficient was 0.82. Other coefficients were also produced. Cronbach's Alpha for Part 1 was 0.58, and for Part 2, 0.63; the correlation between forms was 0.69. Finally, the Guttman Split-Half Coefficient was 0.82. In conclusion, then, the SFT could be considered reliable for this group of students. Perhaps it could also be reliably used with a group of students with similar English proficiency (students who score from approximately 290–500 on the Institutional Pre-TOEFL).

I also examined unidimensionality of SFT test items through inspection of a bubble chart pathway plot produced by Rasch Analysis, using Winsteps. Figure 2 shows the plot; it is a representation of the fit of items of the SFT (specifically looking at infit). According to Bond and Fox (2007), fit statistics help us "to determine whether the item estimations may be held as meaningful quantitative summaries of the observations (i.e., whether each item contributes to the measurement of only one construct)" (p. 35). Acceptable fit on this plot is between −2 and +2 (p. 57). In addition, the size of the circles in this plot is a reflection of error, with larger circles reflecting more error in measurement. In the figure below, the results are generally positive regarding the fit of SFT items because most test items fit within the range of −2 and +2. However, we also can see that some items may need revision; item 10 is overfitting while item 20 is close to underfitting. In addition, items 16 and 18 display relatively more error and should be examined for possible revision as well. In short, despite the need to inspect a few items, this line

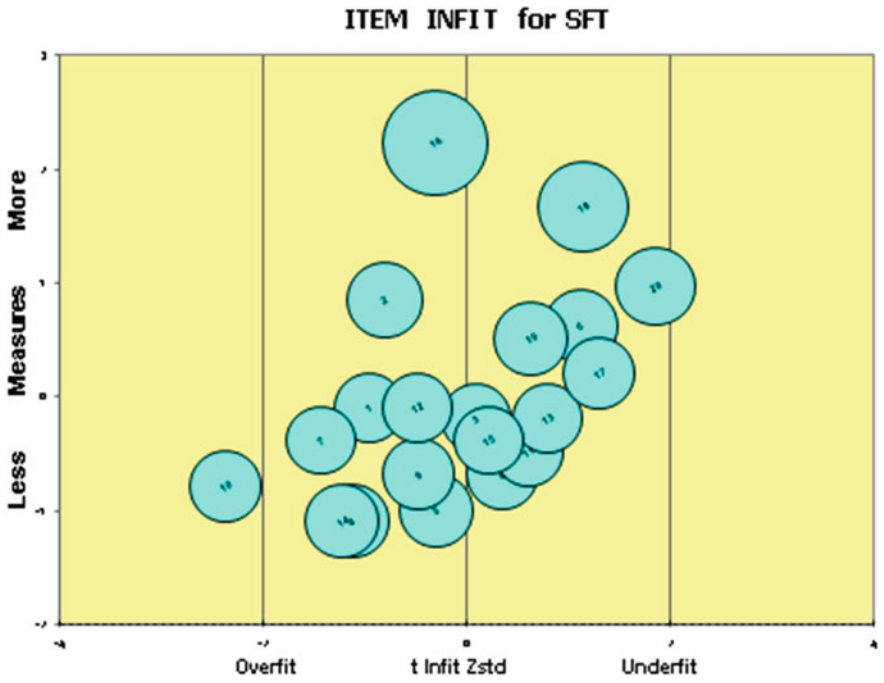


Fig. 2 Item infit for SFT items

of validation evidence for the use of the SFT is mostly positive in that almost all items are contributing in a meaningful manner to measurement of one construct.

The results of Research Question 1 are positive validation evidence in that the test items match student ability, create a useful spread of student ability, generally display acceptable fit and difficulty, and demonstrate acceptable reliability as well as unidimensionality. The test content is achieving its purpose, and generally working in a positive way to measure students in a unidimensional manner.

Research Question 2: Can the SFT be shown to tap into same construct of writing ability that a composition task taps?

In order to answer research question 2, I first examined correlations among all scores. The Pearson product-moment correlation coefficient was calculated for all the comparisons. Specifically, the SFT scores were correlated with the essay scores (produced by two holistic raters and three analytic raters). Naturally, the assumptions underlying the correlation statistic (of independence, normal distribution, and linear relationship) were checked (Brown 1996, p. 157). All assumptions were met. The unadjusted correlations are presented in Table 3. The Bonferroni approach was used to control for Type I error across the 15 correlations. A p value of less than 0.003 (0.05/15 = 0.003) was required for statistical significance (Green and Salkind 2005, p. 261). The results showed that all 15 coefficients were statistically significant and large (Field 2005). Such results suggest that students tended to score in a similar fashion on the SFT and the writing task. As Stansfield and Ross (1988) said, this result suggests that the SFT and writing task both tap into the same construct of writing ability.

I also examined the SFT and essay ratings for unidimensionality using principal components analysis. This type of analysis allows us to examine underlying dimensions, and to determine whether test scores "reflect a single variable" or not (Field 2005, p. 619). The analysis was conducted in order to determine whether the SFT would load together with essay ratings onto the same factor. Table 4 presents the results. The analysis resulted in one component, and an eigenvalue of 3.62, accounting for 72.24 % of the variance. According to Armor (1974), any factor that accounts for 40 to 60 % is a good solution; therefore, these results are favorable. In short, the essay ratings and SFT were fundamentally unidimensional, and seem to

Table 3 Intercorrelations of holistic essay ratings, analytic essay ratings, and SFT scores

Measure	1	2	3	4	5	6
1. HR1	–					
2. HR2	0.83[a]	–				
3. AR1	0.71[a]	0.60[a]	–			
4. AR2	0.69[a]	0.57[a]	0.72[a]	–		
5. AR3	0.70[a]	0.56[a]	0.82[a]	0.82[a]	–	
6. SFT	0.65[a]	0.57[a]	0.63[a]	0.57[a]	0.55[a]	–

Note [a]$p < 0.0001$. $N = 50$ for Essay Ratings and SFT. *HR* Holistic rater; *AR* Analytic rater. *SFT* Sentence form test

Table 4 Factor loadings from principal components analysis of essay ratings and SFT: communalities, eigenvalue, and percentage of variance

Writing measure	Component 1	Communality
HR1	0.90	0.81
HR2	0.81	0.65
AR1	0.88	0.77
AR2	0.86	0.74
AR3	0.88	0.77
SFT	0.77	0.60
% of variance	72.24	

be tapping into the same construct of writing ability. Although a larger sample size would be preferable for principal components analysis, this line of validation evidence also supports using the SFT as a test of writing.

Discussion

As an objective measure of writing ability, the SFT has many obvious advantages for writing class placement. It is administered and scored quickly and easily, and there are no concerns about choice or quality of rubrics, or about rater behavior. Although some may argue the need for "direct" writing measures, because these feel somehow "right," surely professionals must use more than feelings in making testing decisions. Though ratings of writing tasks can work well, as they did in this study, it is clear that good objective measures like the SFT can offer an efficient and reliable supplement to or substitute for traditional rater assessment of writing.

References

Armor, D. J. (1974). Theta reliability and factor scaling. *Sociological Methodology, 5*, 17–50.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, N.J.: Lawrence Erlbaum.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River: Prentice Hall Regents.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*, 21–42.

Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 155–164). Norwood: Ablex Publishing.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

ETS. (2016). Retrieved February 29, 2016, from: https://www.ets.org/toefl/pbt/prepare/sample_questions/structure_written_expression_practice_section2

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

Green, S., & Salkind, N. (2005). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (4th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Hamp-Lyons, L. (1991). Issues and directions in assessing second language writing in academic contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 323–329). Norwood: Ablex Publishing.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley: Newbury House Publishers.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219–239.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's Legacy. *Language Assessment Quarterly, 3*(1), 31–51.

Stansfield, C. W., & Ross, J. (1988). A long-term research agenda for the test of written english. *Language Testing, 5*, 160–186.

Takagi, K. K. (2014, August). *Writing assessment in university entrance examinations: The case of one Japanese university*. Paper presented at the meeting of the Pacific Rim Objective Measurement Symposium, Guangzhou, China.

Developing and Evaluating a Questionnaire to Measure EFL Learners' Vocabulary Learning Motivation

Mitsuko Tanaka

Introduction

Self-determination theory (SDT; Deci and Ryan 1985, 2000, 2002) is one of the most influential motivational theories in the field of educational psychology. Although numerous studies have been conducted to investigate the motivation to learn a second language (L2) using the SDT framework (e.g., Hiromori 2006; Noels et al. 2000; Pae 2008; Tanaka 2013; Vandergrift 2005), there is no research that examines the motivation for vocabulary learning when studying English as a foreign language (EFL) from this perspective. As there is no SDT questionnaire focusing on EFL vocabulary learning motivation, this study aims to develop and evaluate an SDT questionnaire for EFL vocabulary learning using Rasch analysis.

Self-determination Theory

SDT (Deci and Ryan 2002) categorizes motivation into three broad categories: intrinsic motivation, extrinsic motivation, and amotivation. Intrinsic motivation refers to motivation to engage in an activity for the sake of one's own enjoyment. Extrinsic motivation refers to motivation driven by external rewards. Amotivation is a state of lack of motivation. Extrinsic motivation is further classified into four types of regulation, three of which (identified, introjected, and external regulation) have been employed in empirical studies in L2 motivational literature (e.g., Noels et al. 2000; Tanaka 2013). Identified regulation is a state regulated by the importance and

M. Tanaka (✉)
Ritsumeikan University, Shiga, Japan
e-mail: mtanaka@fc.ritsumei.ac.jp

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_23

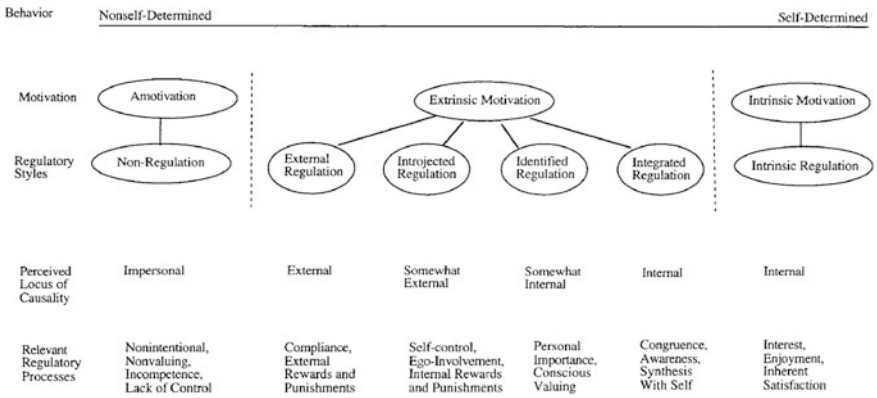


Fig. 1 The self-determination continuum (Ryan and Deci 2000, p. 72)

values of learning. For example, learners with identified regulation study English vocabulary because they believe that English vocabulary is useful or important to accomplish their life goals. Introjected regulation concerns the maintenance of a person's self-worth. For instance, students study English vocabulary because they do not want their classmates to think that they are poor at English or slow in acquiring English vocabulary. External regulation is a state regulated by rewards or punishments. For example, learners with external regulation study English vocabulary because they want to get course credits, grades, or high test scores. According to SDT, these five types of motivation and regulation are ordered from intrinsic motivation to amotivation on a continuum (Fig. 1) and have a simplex-like structure. Theoretically, adjunct regulations on the continuum should be correlated more highly than regulations situated further apart.

Research Purposes

As discussed above, there is no questionnaire for EFL vocabulary learning motivation using the SDT framework. This study aims to develop and evaluate an SDT questionnaire for EFL vocabulary learning using Rasch analysis.

1. Does each item function properly?
2. Is each of the five constructs reliable?
3. Is each of the five constructs unidimensional?
4. Is the 6-point rating scale psychometrically optimal?
5. Do the five constructs form the simplex-like structure that SDT postulates?

Method

Creation of a Questionnaire for SDT Vocabulary Learning Motivation

An SDT questionnaire for English vocabulary learning motivation was developed drawing primarily from Tanaka (2013, 2014). The developed questionnaire consists of five constructs (intrinsic motivation, identified regulation, introjected regulation, external regulation, and amotivation for learning English vocabulary), with five items in each construct. The questionnaire is a 6-point Likert scale ranging from 1 (Strongly disagree) to 6 (Strongly agree). See Appendix for the English translation of the questionnaire.

Data and Sample

The data for this study comes from first-year science and engineering students (N = 179; mostly male students aged between 15 and 16) at a public technical college in Japan. They took five English classes per week (45 min per session) and vocabulary lists were assigned as weekly homework over the year. At the time of data collection, the students had completed approximately four years of compulsory English learning at secondary schools (i.e., at junior high school and college). The questionnaire was administered in Japanese in classes around the end of the 2012 academic year.

Data Analysis Procedures

Rasch analyses were performed using Winsteps 3.80.0 as follows. First, a Rasch fit analysis was conducted to examine items and the reliability of each construct measured by the questionnaire. Second, a Rasch principal components analysis (PCA) of item residuals was conducted to examine dimensionality of each construct. Third, the rating scale categories of each construct were assessed and optimized based on Linacre's (2002) six criteria. In addition to these Rasch analyses, Pearson product-moment correlation coefficients were calculated among the five constructs using SPSS 19.0 to examine the simplex-like structure that SDT postulates.

Results

Items

First, a Rasch fit analysis was conducted to examine items designed to measure each construct in the questionnaire. The criteria for acceptable items were the infit and outfit mean square (MNSQ) statistics of 0.50–1.50 (Linacre 2012, p. 553). Table 1 shows the summary of the Rasch item fit statistics. All infit and outfit MNSQ statistics were between 0.50 and 1.50 (Max. = 1.38, Min. = 0.68 for infit MNSQ statistics; Max. = 1.48, Min. = 0.69 for outfit MNSQ statistics). The point-measure correlations of the items were adequately high ($M = 0.82$, Max. = 0.91, Min. = 0.75). Taken together, each item functioned properly and adequately contributed to measuring the intended construct.

Reliability and Separation of Measures

The reliability of each construct was examined based on Rasch person reliability and separation estimates and Rasch item reliability and separation estimates. The criteria for person estimates are above 0.80 for reliability and above 2.0 for separation, as person reliability of 0.80 indicates the presence of 2 or 3 statistically distinct levels in the sample (Linacre 2012, p. 574). As shown in Table 2, while three constructs (intrinsic motivation, introjected regulation, and amotivation) satisfied the criteria, two constructs (identified and external regulation) showed person reliability and separation estimates slightly lower than the criteria (reliability: 0.7, separation: 1.81 and 1.82). Consequently, eight misfitting people (5 % of the total of 179 participants) were temporarily eliminated based on the analysis of the most unexpected responses, and person reliability (separation) were recalculated for identified and external regulation. As a result, person reliability (separation) improved into 0.82 (2.12) for identified regulation and 0.81 (2.05) for external regulation, satisfying the criteria. Given that the elimination of a very small number of misfitting people improved the reliability (separation) estimates, person reliability of each construct was adequately high.

With respect to item estimates, a reliability estimate above 0.90 and a separation above 3.0 are considered ideal values, as this confirms the item difficulty hierarchy (low, medium, and high difficulties) of the instrument (Linacre 2012, p. 575). Most of the item reliability (separation) estimates were very high, being above or very close to the ideal values of 0.90 (3.00). However, one construct (external regulation) showed low item reliability (0.67) and separation (1.47) estimates. When the eight misfitting people were temporality eliminated, item reliability (separation) improved to 0.77 (1.82). Although these values are still lower than the ideal values, the reliability of 0.77 is not considered very problematic as it is close enough to the value of 0.80 where items are stratified between 2 and 3 levels in terms of difficulty.

Table 1 Rasch item fit statistics for the five SDT motivation items

Item	Measure	*SE*	Infit	Infit	Outfit	Outfit	PMC
			MNSQ	ZSTD	MNSQ	ZSTD	
Intrinsic motivation for learning English Vocabulary (IM)							
IM1	0.49	0.11	0.98	−0.17	0.94	−0.45	0.82
IM2	0.73	0.12	0.75	−2.35	0.69	−2.56	0.84
IM3	−0.90	0.11	1.38	3.02	1.48	3.51	0.79
IM4	−0.91	0.11	1.01	0.09	1.00	0.01	0.83
IM5	0.59	0.12	0.82	−1.59	0.81	−1.57	0.83
Identified regulation for learning English Vocabulary (ID)							
ID1	−0.49	0.12	1.02	0.21	1.04	0.43	0.76
ID2	0.62	0.12	1.25	2.11	1.24	1.87	0.75
ID3	−0.49	0.12	0.84	−1.51	0.80	−1.86	0.80
ID4	0.57	0.12	0.86	−1.27	0.80	−1.68	0.81
ID5	−0.21	0.11	1.03	0.28	1.09	0.83	0.75
Introjected regulation for learning English Vocabulary (IJ)							
IJ1	−0.71	0.12	1.32	2.51	1.29	2.28	0.88
IJ2	0.04	0.12	1.07	0.65	1.09	0.79	0.88
IJ3	0.21	0.12	0.68	−3.02	0.70	−2.75	0.91
IJ4	0.12	0.12	1.01	0.14	0.98	−0.10	0.88
IJ5	0.34	0.12	0.86	−1.24	0.90	−0.82	0.89
External regulation for learning English Vocabulary (EX)							
EX1	0.07	0.12	1.02	0.24	0.99	−0.05	0.79
EX2	0.22	0.12	0.78	−2.10	0.81	−1.73	0.83
EX3	0.19	0.12	1.03	0.34	0.99	−0.05	0.79
EX4	−0.29	0.12	1.09	0.88	1.09	0.84	0.77
EX5	−0.19	0.12	1.06	0.61	1.04	0.35	0.78
Amotivation for learning English Vocabulary (AM)							
AM1	0.18	0.13	1.11	1.00	1.12	1.04	0.80
AM2	−0.80	0.13	1.00	0.03	0.99	0.00	0.85
AM3	0.61	0.13	1.03	0.31	1.02	0.21	0.79
AM4	−0.29	0.13	0.82	−1.65	0.84	−1.38	0.85
AM5	0.30	0.13	1.05	0.48	1.05	0.43	0.80
M	0.00	0.12	1.00	−0.08	0.99	−0.10	0.82
Max.	0.73	0.13	1.38	3.02	1.48	3.51	0.91
Min.	−0.91	0.11	0.68	−3.02	0.69	−2.75	0.75

Note: *PMC* Point-measure correlation

However, some improvement in reliability and separation is recommended for a revised version. As low item reliability indicates "a narrow range of item measures, or a small sample" (p. 575), a revised version should have more items with a wider difficulty range or should be tested with a larger sample with wider ability variance.

Table 2 Reliability and separation of the five constructs

Components checked	Values				
	IM	ID	IJ	EX	AM
Item separation	6.23	4.00	2.68	1.47 (1.82)	3.60
Item reliability	0.97	0.94	0.88	0.67 (0.77)	0.93
Person separation	2.07	1.82 (2.12)	2.69	1.81 (2.05)	2.00
Person reliability	0.81	0.77 (0.82)	0.88	0.77 (0.81)	0.80

Note Values within the brackets represent estimates when the eight misfitting people are removed *IM* Intrinsic Motivation; *ID* Identified Regulation; *IJ* Introjected Regulation; *EX* External Regulation; *AM* Amotivation

Dimensionality

Dimensionality of each construct was examined using the Rasch PCA of item residuals. Construct unidimensionality is assessed in terms of variance explained and variance unexplained by the Rasch measures. Table 3 shows the results of the analysis. The four constructs (intrinsic motivation, identified regulation, introjected regulation, and amotivation) had an adequate amount of variance explained by the Rasch measures as they were more than the half the total variance. Concerning unexplained variance, the ideal eigenvalue of the first contrast in the residuals was less than 2.0 (Linacre 2012, p. 353). In practice, however, the eigenvalue should be less than 3.0, as the strength of at least 3 items (i.e., eigenvalue of 3.0) is necessary to form a secondary dimension (p. 496). As shown in Table 3, all the five constructs had eigenvalues less than 3.0 and thus satisfied the practical criterion. However, the unexplained variance in percentages appeared to be somewhat large. In particular, external regulation had a large amount of residuals for the first contrast (30 %), which was greater than the variance explained by the item difficulties (20.1 %). The total unexplained variance (58.1 %) was also larger than the total variance

Table 3 Rasch PCA of item residuals of the five constructs

Variance component		Values				
		IM	ID	IJ	EX	AM
Raw variance explained by measures		65.60	51.50	66.30	41.90	60.50
	Raw variance explained by persons	40.90	28.4	50.9	21.8	41.6
	(eigenvalue)	6.0	2.9	7.6	1.9	5.3
	Raw variance explained by items	24.7	23.1	15.4	20.1	18.9
	(eigenvalue)	3.6	2.4	2.3	1.7	2.4
Raw unexplained variance		34.40	48.5	33.7	58.1	39.5
	Unexplained variance in 1st contrast	16.80	19.60	10.90	30.00	11.90
	(eigenvalue)	2.50	2.00	1.60	2.60	1.50

Note: *IM* Intrinsic Motivation; *ID* Identified Regulation; *IJ* Introjected Regulation; *EX* External Regulation; *AM* Amotivation

Table 4 Positively and negatively loading items in the Rasch PCA of item residuals for external regulation

Items	Loading	Measure	Infit MNSQ	Outfit MNSQ	Item Description
EX4	0.85	−0.29	1.09	1.09	Because I want to get English course credits
EX5	0.85	−0.19	1.06	1.04	Because I don't want to fail the English course
EX2	−0.68	0.22	0.78	0.81	Because I want to get good grades
EX1	−0.62	0.07	1.02	0.99	Because I want to get high scores on tests
EX3	−0.54	0.19	1.03	0.99	Because I don't want to get bad grades

explained by measures (41.90 %). Given that the total explained variance should ideally be four times larger than the total unexplained variance (p. 496), residuals in the construct of external regulation is very large. As such, item loadings were examined to explore a possible secondary dimension.

As shown in Table 4, the five items were separated into two clusters. Whereas two items with high positive loadings (EX4 and EX5) concern course credits, three items with high negative loadings (EX1, EX2, and EX3) are grade-related items. The disattentuated person measures from these two clusters of items showed a mere medium correlation ($r = 0.53$, $p < 0.001$).

Although some degree of multidimensionality is suggested for this construct, "[m]ultidimensionality always exists to a lesser or greater extent" (Linacre 2012, p. 497). Examination of the content of items is also important to determine unidimensionality. Linacre (p. 489) suggested the following guidelines for determining unidimensionality:

> [L]ook at the content (wording) of the items. If those items are different enough to be considered different dimensions (similar to "height" and "weight"), then split the items into separate analyses. If the items are part of the same dimension (similar to "addition" and "subtraction" on an arithmetic test), then no action is necessary. You are seeing the expected co-variance of items in the same content area of a dimension.

The theoretical content of external regulation represents a "broad" motivational state regulated by external factors such as rewards and punishment, which include grades, scores, and credits. Although the removal of either grade- or credit-related items improves the Rasch unidimensionality of this construct, both clusters are part of the same external regulation. As such, it is not necessary to separate the items into two constructs.

Rating Scale Categories

The effectiveness of the original 6-point rating scale categories (1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, and 6 = *Strongly agree*) was examined and optimized based on Linacre's (2002) six guidelines:

1. Each category should have more than 10 observations;
2. Each category should have a peak in the probability curve;
3. The average category measures should progress with the rating scale categories;
4. Outfit mean squares should be smaller than 2.0;
5. Threshold calibration should progress with the rating scale category; and
6. The category threshold should be between 1.4 and 5.0 logits apart.

Concerning the sixth criterion, the minimum threshold separation was assessed based on Wolfe and Smith's (2007) criteria: 0.59, 0.81, 1.1, and 1.4 for a 6-, 5-, 4-, and 3-point scale, respectively. When the above six criteria were not satisfied, the rating scale categories were optimized by combining categories.

Table 5 shows the summary of the category structure for intrinsic motivation. In the 6-point rating scale, the separation between the first and second thresholds ($\tau_1 = -2.01$, $\tau_2 = -1.71$) was 0.30, which was well below the required 0.59 logits

Table 5 Summary of the category structure for intrinsic motivation

Category label	Count	(%)	Average measure	Infit MNSQ	Outfit MNSQ	Structure measure
The 6-point rating scale						
1 Strongly disagree	184	(21)	−2.60	1.09	1.15	NONE
2 Disagree	120	(14)	−1.69	0.73	0.69	**−2.01**
3 Slightly disagree	201	(23)	−0.81	0.78	0.70	**−1.71**
4 Slightly agree	212	(24)	0.37	0.76	0.77	−0.35
5 Agree	97	(11)	1.40	1.07	1.23	1.53
6 Strongly agree	72	(8)	2.40	1.77	1.60	2.55
The 5-point rating scale						
1 Disagree	304	(34)	−2.70	1.03	1.03	NONE
2 Slightly disagree	201	(23)	−1.62	0.82	0.81	−2.12
3 Slightly agree	212	(24)	−0.21	0.78	0.70	−1.03
4 Agree	97	(11)	0.92	1.09	1.24	1.01
5 Strongly agree	72	(8)	2.08	1.49	1.46	2.13

Note Boldface indicates values that did not meet the criteria

for a 6-point rating scale. In the 5-point rating scale, when categories 1 and 2 were combined, the separation between the thresholds became 1.09 ($\tau_1 = -2.12$, $\tau_2 = -1.03$), which was greater than the required 0.81 for a 5-point rating scale. The other criteria were also satisfied as all the categories had more than 10 observations; the outfit mean square statistics were below 2.0; the average category measures were ordered, progressing from −2.70 for category 1 to 2.08 for category 5; the shape of the probability curves was peaked for each category (Fig. 2). Thus, the 5-point rating scale was considered optimal for the construct of intrinsic motivation.

Table 6 shows the summary of the category structure for identified regulation. In the 6-point rating scale, threshold measures were disordered between categories 1 and 2. In the 5-point rating scale when these categories are combined, the threshold measures were ordered but the separation between the thresholds (0.35, $\tau_1 = -1.40$, $\tau_2 = -1.05$) was well below the required 0.81 for a 5-point rating scale. Consequently, categories 1 and 2 were combined again. The separation between the first and second thresholds ($\tau_1 = -1.49$, $\tau_2 = -0.12$) became 1.37, which was larger than the required 1.1 for 4-point rating scale. The other criteria were also satisfied (see Table 6 and Fig. 3). Thus, the 4-point rating scale was considered optimal for the construct of identified regulation.

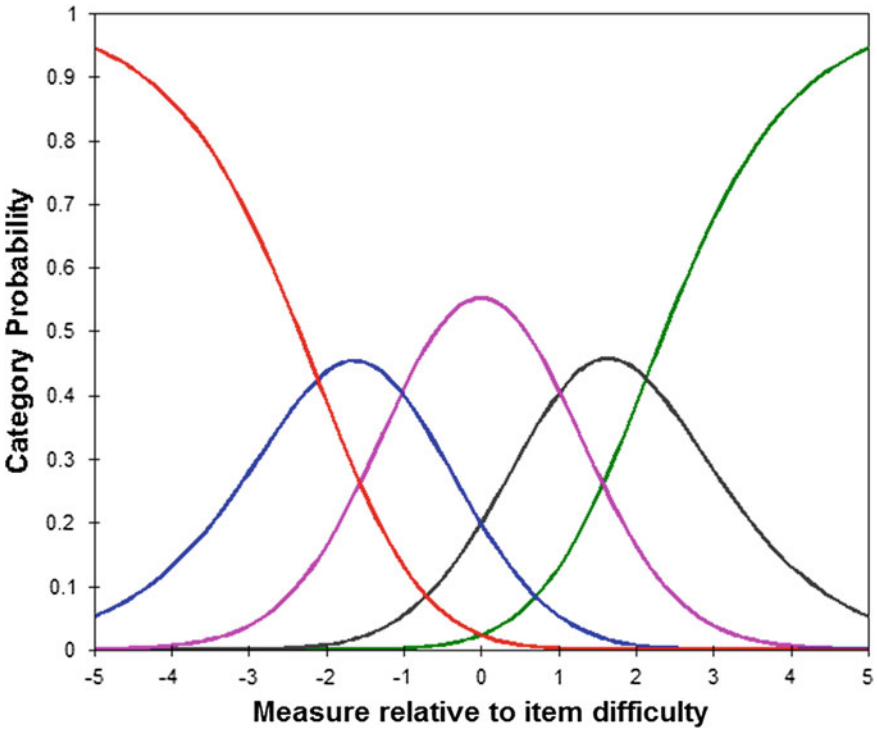


Fig. 2 The 5-point rating scale performance for intrinsic motivation

Table 6 Summary of the category structure for identified regulation

Category label	Count	(%)	Average measure	Infit MNSQ	Outfit MNSQ	Structure measure
The 6-point rating scale						
1 Strongly disagree	74	(8)	−0.98	1.30	1.42	NONE
2 Disagree	45	(5)	−0.83	0.75	0.68	**−0.82**
3 Slightly disagree	126	(14)	−0.42	0.68	0.63	**−1.48**
4 Slightly agree	247	(28)	0.31	0.74	0.84	−0.66
5 Agree	228	(26)	1.17	0.70	0.73	0.70
6 Strongly agree	168	(19)	1.76	1.71	1.32	2.26
The 5-point rating scale						
1 Disagree	119	(13)	−1.32	1.22	1.22	NONE
2 Slightly disagree	126	(14)	−0.92	0.68	0.67	**−1.40**
3 Slightly agree	247	(28)	−0.04	0.82	0.87	**−1.05**
4 Agree	228	(26)	0.91	0.72	0.76	0.42
5 Strongly agree	168	(19)	1.61	1.45	1.30	2.03
The 4-point rating scale						
1 Disagree	245	(28)	−1.74	1.09	1.07	NONE
2 Slightly disagree	247	(28)	−0.78	0.87	0.91	−1.49
3 Slightly agree	228	(26)	0.42	0.78	0.82	−0.12
4 Agree	168	(19)	1.30	1.20	1.19	1.62

Note Boldface indicates values that did not meet the criteria

Table 7 shows the summary of the category structure for introjected regulation. All the categories had more than 10 observations; the outfit mean square statistics were below 2.0; the average category measures were ordered, progressing from −3.47 for category 1–3.19 for category 6, and the smallest separation between the thresholds was 1.71 ($\tau_5 = 2.17$, $\tau_6 = 3.88$), which was greater than the required 0.59 logits for a 6-point rating scale. Moreover, the shape of the probability curves peaked for each category (Fig. 4). Thus, the 6-point rating scale was optimal for this construct.

Table 8 shows the summary of the category structure for external regulation. Categories 1 and 2 were combined twice, as the separation between the first and the

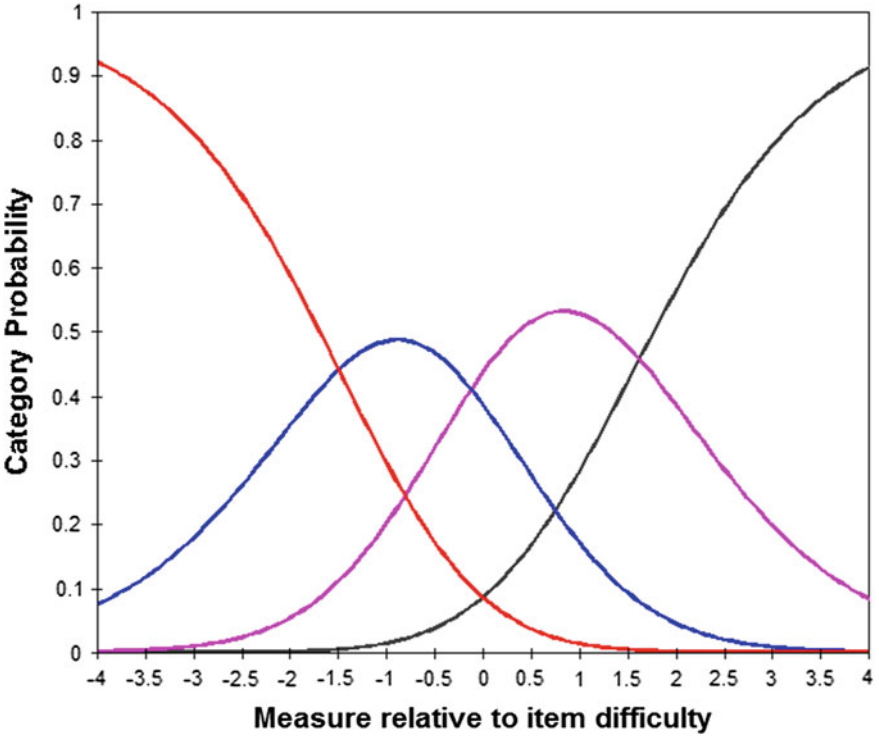


Fig. 3 The 4-point rating scale performance for identified regulation

Table 7 Summary of the category structure for introjected regulation

Category label	Count	(%)	Average measure	Infit MNSQ	Outfit MNSQ	Structure measure
The 6-point rating scale						
1 Strongly disagree	222	(25)	−3.47	1.56	1.42	NONE
2 Disagree	195	(22)	−2.51	0.62	0.64	−4.28
3 Slightly disagree	211	(24)	−0.87	0.74	0.72	−1.75
4 Slightly agree	167	(19)	0.57	0.91	0.97	−0.02
5 Agree	62	(7)	1.66	1.48	1.53	2.17
6 Strongly agree	31	(3)	3.19	1.17	1.15	3.88

Note Boldface indicates values that did not meet the criteria

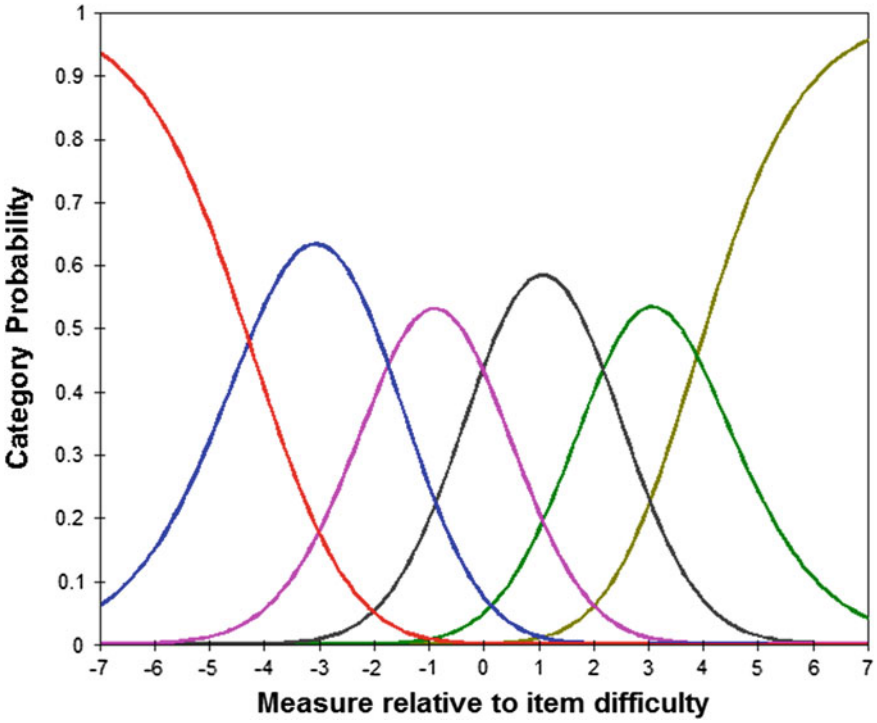


Fig. 4 The 6-point rating scale performance for introjected regulation

second thresholds was well below the required value of 0.59 logits for a 6-point rating scale and 0.81 logits for a 5-point rating scale. In the 4-point rating scale, the smallest separation between the thresholds (1.42, $\tau_3 = 0.08$, $\tau_4 = 1.46$) was greater than the required 1.1 for a 4-point rating scale. The other criteria were also satisfied (see Table 8 and Fig. 5). Thus, the 4-point rating scale was considered optimal for the construct of external regulation.

Table 9 shows a summary of the category structure for amotivation. Categories 5 and 6 in the 6-point and categories 4 and 5 in the 5-point rating scales were combined as threshold measures were reversed. In the 4-point rating scale, the threshold measures were ordered. The other criteria were also satisfied (see Table 9 and Fig. 6). Thus, the 4-point rating scale was considered optimal for the construct of identified regulation.

Table 10 shows the results of rating scale optimization. The 6-point rating scale was retained only for introjected regulation. Scales were reduced into 5-point rating scales for intrinsic motivation, and 4-point rating scales for identified regulation, external regulation, and amotivation.

Table 8 Summary of the category structure for external regulation

Category label	Count	(%)	Average measure	Infit MNSQ	Outfit MNSQ	Structure measure
The 6-point rating scale						
1 Strongly disagree	55	(6)	−1.09	1.41	1.74	NONE
2 Disagree	50	(6)	−0.69	0.98	1.01	**−1.31**
3 Slightly disagree	103	(12)	−0.18	0.84	0.87	**−1.06**
4 Slightly agree	259	(29)	0.40	0.77	0.78	**−0.74**
5 Agree	213	(24)	1.14	0.72	0.77	0.93
6 Strongly agree	212	(24)	1.61	1.23	1.11	2.19
The 5-point rating scale						
1 Disagree	105	(12)	−0.91	1.30	1.38	NONE
2 Slightly disagree	103	(12)	−0.66	0.88	0.92	**−1.27**
3 Slightly agree	259	(29)	0.00	0.80	0.81	**−1.16**
4 Agree	213	(24)	0.80	0.74	0.76	0.57
5 Strongly agree	212	(24)	1.32	1.13	1.09	1.86
The 4-point rating scale						
1 Disagree	208	(23)	−1.27	1.27	1.25	NONE
2 Slightly disagree	259	(29)	−0.66	0.81	0.83	−1.54
3 Slightly agree	213	(24)	0.35	0.78	0.74	0.08
4 Agree	212	(24)	0.99	1.01	1.01	1.46

Note Boldface indicates values that did not meet the criteria

The Theoretical Tenets of the Simplex-Like Structure of the SDT Scale

As discussed earlier, SDT (Deci and Ryan 2002) postulates a simplex-like pattern on the continuum of the five subscales, where adjunct regulations have a stronger and positive correlation with each other. The results of the correlation analysis showed that the five constructs have the simplex-like structure that SDT postulates (Table 11). As such, the measurement of the five constructs adequately represents SDT.

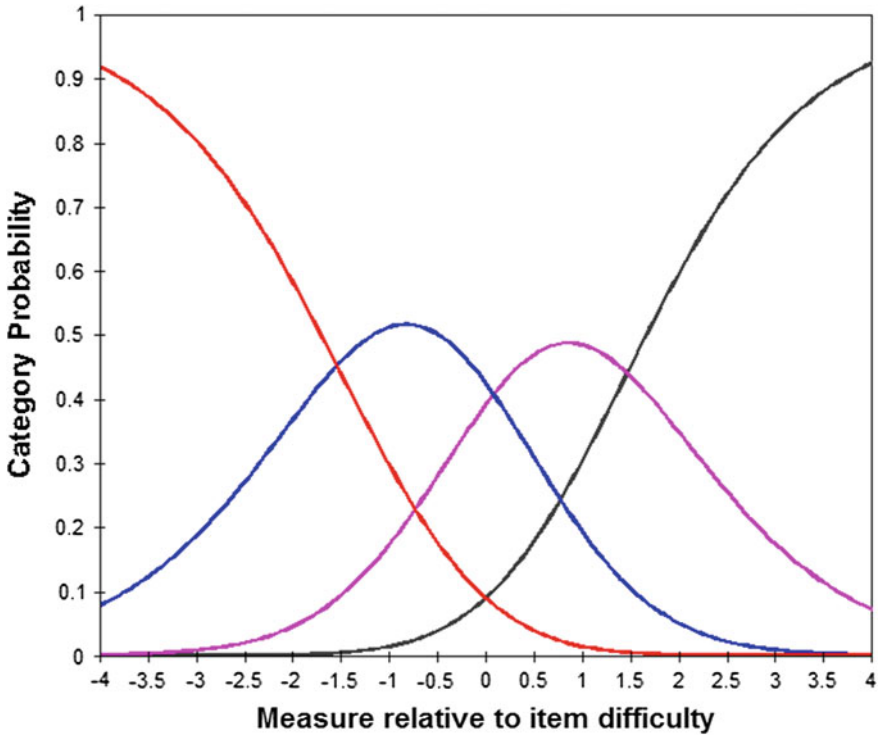


Fig. 5 The 4-point rating scale performance for external regulation

Table 9 Summary of the category structure for amotivation

Category label	Count	(%)	Average measure	Infit MNSQ	Outfit MNSQ	Structure measure
The 6-point rating scale						
1 Strongly disagree	269	(30)	−2.73	1.42	1.11	NONE
2 Disagree	264	(30)	−1.62	0.66	0.74	−2.69
3 Slightly disagree	213	(24)	−0.32	0.69	0.71	−0.72
4 Slightly agree	77	(9)	0.45	0.76	0.73	0.97
5 Agree	34	(4)	0.84	0.87	0.87	**1.37**
6 Strongly agree	32	(4)	0.89	1.54	2.62	**1.06**

(continued)

Table 9 (continued)

Category label	Count	(%)	Average measure	Infit MNSQ	Outfit MNSQ	Structure measure
The 5-point rating scale						
1 Strongly disagree	269	(30)	−2.54	1.27	1.11	NONE
2 Disagree	264	(30)	−1.32	0.70	0.76	−2.46
3 Slightly disagree	213	(24)	0.09	0.73	0.72	−0.39
4 Slightly agree	77	(9)	1.02	0.82	0.83	**1.47**
5 Agree	66	(7)	1.35	1.39	1.80	**1.38**
The 4-point rating scale						
1 Strongly disagree	269	(30)	−2.23	1.13	1.12	NONE
2 Disagree	264	(30)	−0.84	0.75	0.75	−2.08
3 Slightly agree	213	(24)	0.84	0.78	0.79	0.18
4 Agree	143	(16)	2.05	1.34	1.44	1.91

Note Boldface indicates values that did not meet the criteria

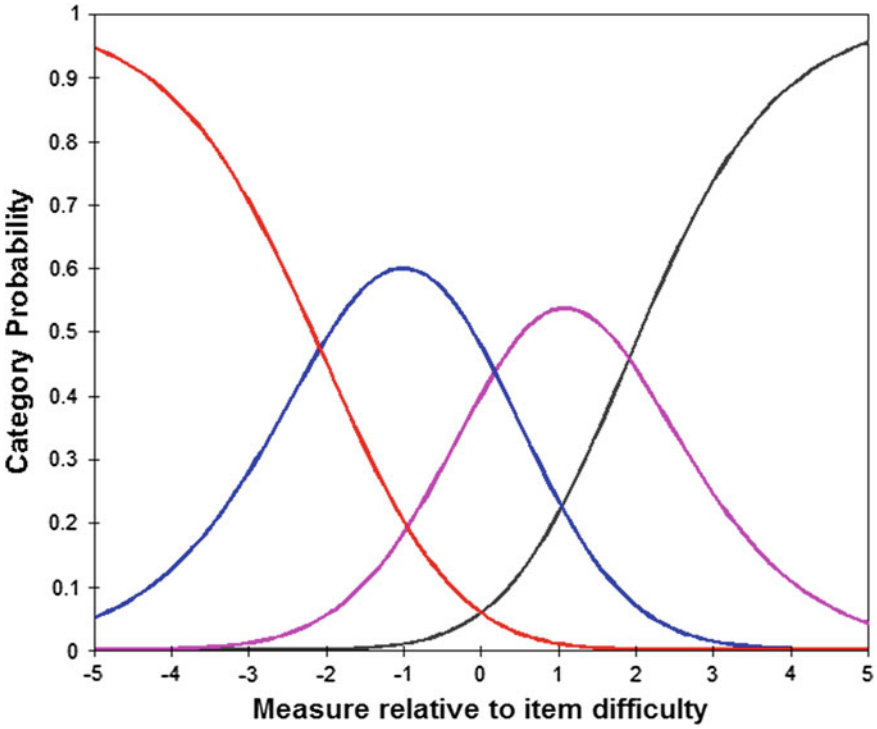


Fig. 6 The 4-point rating scale performance for amotivation

Table 10 Summary of the rating scale optimization

Constructs	The resulting rating scale
Intrinsic motivation	5-point scale (112345)
Identified regulation	4-point scale (111234)
Introjected regulation	6-point scale (123456)
External regulation	4-point scale (111234)
Amotivation	4-point scale (123444)

Table 11 Correlation matrix of the five constructs

Factor	Intrinsic motivation	Identified regulation	Introjected regulation	External regulation
Identified regulation	0.57	–		
Introjected regulation	0.22	0.13	–	
External regulation	0.02	0.15	0.14	–
Amotivation	−0.33	0.52	0.26	−0.12

Conclusion

The present study aimed to develop and evaluate an SDT questionnaire for EFL vocabulary learning motivation using Rasch analysis. First, the results of a Rasch fit analysis showed that each item functions properly and adequately contributes to measuring the intended construct. Second, the results of Rasch reliability and separation analyses revealed that both person and item reliability (separation) were adequately high, although some improvement was recommended for external regulation. Third, the results of the Rasch PCA of item residuals showed that constructs were adequately unidimensional. Fourth, the results of rating scale analysis showed that the original 6-point rating scale was retained only for introjected regulation. The rating scale categories of the remaining four constructs were properly optimized by reducing categories. Fifth, the results of the correlation analysis showed that the measurement of the five constructs adequately represented the self-determination theory. Taken together, the developed SDT questionnaire instrument for EFL vocabulary learning motivation was adequately valid and reliable for the participants of the present study.

Appendix: Questionnaire Items (English Translation)

Why do you study English vocabulary?

Factor 1: Intrinsic Motivation for Learning English Vocabulary (IM)	
IM1	Because learning English vocabulary is enjoyable
IM2	Because learning English vocabulary is interesting
IM3	Because I feel pleasure when I discover new things through learning English vocabulary
IM4	Because I feel pleasure about increasing my English vocabulary
IM5	Because I like learning English vocabulary
Factor 2: Identified Regulation for Learning English Vocabulary (ID)	
ID1	Because English vocabulary is useful
ID2	Because English vocabulary is important to make my dreams come true
ID3	Because English vocabulary will be necessary in the future
ID4	Because English vocabulary is necessary to attain my life goals
ID5	Because it is important to acquire English vocabulary
Factor 3: Introjected Regulation for Learning English Vocabulary (IJ)	
IJ1	Because I'd feel ashamed if I have smaller amount of English vocabulary than my classmates
IJ2	Because I'd feel ashamed if my classmates think that I am an incapable student
IJ3	Because I don't want my classmates to think that I am poor at English
IJ4	Because I don't want my classmates to think that I don't have an adequate amount of English vocabulary
IJ5	Because I don't want my classmates to think that I am slow in acquiring English vocabulary compared to others
Factor 4: External Regulation for Learning English Vocabulary (EX)	
EX1	Because I want to get high scores on tests
EX2	Because I want to get good grades
EX3	Because I don't want to get bad grades
EX4	Because I want to get English course credits
EX5	Because I don't want to fail the English course
Factor 5: Amotivation for Learning English Vocabulary (AM)	
AM1	I won't get anything out of learning English vocabulary
AM2	I don't know what I am getting out of learning English vocabulary
AM3	Learning English vocabulary is useless
AM4	I cannot see why I have to study English vocabulary
AM5	Learning English vocabulary is meaningless

Note All the questionnaire items are randomly ordered 6-point Likert scale items

References

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*, 227–268.

Deci, E. L., & Ryan, R. M. (Eds.). (2002). *Handbook of self-determination research*. Rochester, NY: University of Rochester Press.

Hiromori, T. (2006). *Theories and practices that increase the motivation of foreign language learners*. Tokyo: Taga Shuppan.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85–106.

Linacre, J. M. (2012). *A user's guide to WINSTEPS: Rasch-model computer program*. Chicago, IL: MESA.

Noels, K. A., Pelletier, L., Clément, R., & Vallerand, R. (2000). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning, 50*, 57–85.

Pae, T. (2008). Second language orientation and self-determination theory: A structural analysis of the factors affecting second language achievement. *Journal of Language and Social Psychology, 27*, 5–27.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78.

Tanaka, M. (2013). Examining kanji learning motivation using self-determination theory. *System, 41*, 804–816.

Tanaka, M. (2014). The effects of affective variables and kanji growth on L1 Chinese JSL learners' kanji learning (Doctoral dissertation, 2014). *Available from ProQuest Dissertations and Theses database*. (UMI No. 3611180).

Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics, 26*, 70–89.

Wolfe, E. W., & Smith, E. V, Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II–validation activities. *Journal of applied measurement, 8*, 204–234.

Using Person Fit and Person Response Functions to Examine the Validity of Person Scores in Computer Adaptive Tests

A. Adrienne Walker and George Engelhard Jr.

Treating the inferences from all test scores as if they are equally trustworthy is problematic. Test score inferences from persons who do not exhibit adequate model-data fit may not provide accurate inferences about their knowledge, skills, or abilities. Aggregate level person fit analyses provide validity evidence that can be used to inform test score interpretation and use in general. Individual person fit procedures provide further validity evidence that can be used to inform test score interpretation for specific test-takers.

In addition to providing valuable validity information, individual person fit analyses are important for testing practice because most test accountability stakes occur at the individual test-taker level. For computer adaptive tests (CAT), the information obtained from an individual person fit analysis is even more relevant. Traditional item quality-checking and aggregate person quality-checking procedures have restricted utility in CAT because each test-taker can potentially receive a different set of items. Individual person fit analysis in CAT can provide a customized quantification of how well a person's responses accord with the model used to generate his or her achievement level. With increases in the numbers of CAT being administered (Chang and Ying 2009), it is important to study procedures that can provide additional and post-test validity evidence.

The purpose of this study is to explore person fit in a computer adaptive test using a two-step procedure that incorporates statistical and graphical techniques. First, person fit statistics were used to statistically quantify misfit. Then, person response functions (PRF, Trabin and Weiss 1979) were used to graphically depict misfit. This general approach to examining fit has been explored using paper–pencil

A.A. Walker (✉)
Division of Educational Studies, Emory University,
North Decatur Building, Suite 240, Atlanta, GA 30322, USA
e-mail: angela.adrienne.walker@emory.edu

G. Engelhard Jr.
The University of Georgia, Athens, GA, USA

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_24

tests (Emons et al. 2005; Nering and Meijer 1998; Perkins et al. 2011; Ferrando 2014; Walker et al. 2016) and it seems extendable and useful for exploring person fit in CAT. The research question that guides this study is: Do person response functions in conjunction with person fit statistics have the potential to detect and inform researchers of misfit in CAT?

Background

Many methods exist for examining person fit in the context of paper-pencil tests (Karabatsos 2003; Meijer and Sijtsma 2001). By contrast, the research examining person fit in CAT is sparse (Meijer and van Krimpen-Stoop 2010; van Krimpen-Stoop and Meijer 1999, 2000). Much of the research conducted on person fit in CAT uses traditional paper–pencil person fit statistics for detecting person misfit. Researchers have reported that the sampling distributions of some traditional person fit statistics do not hold to their theoretical distributions in CAT, and this makes the interpretation of misfit difficult (Glas et al. 1998; McLeod and Lewis 1999; Nering 1997; van Krimpen-Stoop and Meijer 1999).

Some researchers such as McLeod and Lewis (1999), Meijer (2005), and van Krimpen-Stoop and Meijer (2000), have proposed and evaluated adaptive test-specific person fit statistics. The results have been mixed. Meijer (2005) reported that the detection power of an adaptive test-specific person fit statistic was higher than the detection power of other person fit methods in CAT, which included traditional person fit statistics. van Krimpen-Stoop and Meijer (2000) reported similar detection rates for their adaptive test-specific person fit statistic as was found for traditional person fit statistics in paper-pencil tests, but McLeod and Lewis (1999) reported that their adaptive test-specific person fit statistic was not powerful for detecting misfit in an adaptive test.

Some of these same researchers have promoted using a different statistical framework for conceptualizing person fit in CAT because the items on each adaptive test cover a different range of difficulty. Bradlow et al. (1998) and van Krimpen-Stoop and Meijer (Meijer and van Krimpen-Stoop 2010; van Krimpen-Stoop and Meijer 2000, 2001) introduced the cumulative sum procedure, CUSUM, for detecting person misfit in CAT. Both sets of researchers argue that the CUSUM procedure is useful for person misfit detection in CAT.

In summary, previous research suggests that procedures used to detect person misfit in CAT are not well-understood. An approach that provides researchers with more than one piece of person fit information may be needed to best understand person *mis*fit in CAT. The two-step approach in this study uses both statistical and graphical information to examine and illustrate person fit.

Theoretical Framework

In computer adaptive testing, tests are comprised of items that are individually selected to target the achievement for each test-taker by using a series of rule-based algorithms. These algorithms are heavily reliant on item response theory (IRT) to implement, and the banked item parameters are considered to be fixed and known. The Rasch model (Rasch 1960/1980) is theoretically compatible with an adaptive test procedure because person estimates from a Rasch-calibrated item bank are statistically equivalent across all customized tests, regardless of difficulty (Bond and Fox 2015).

Responses to adaptive test items are stochastic, but a person's responses should still accord with the model chosen to calibrate the items and generate the final score. In other words, adequate person fit should be observed. Person response functions show the relationship between the person's probability of giving the correct response and the difficulty of the items to which she or he responds (Trabin and Weiss 1979). Rasch-based PRF provide a clear graphical illustration of what good person fit looks like. By visually comparing this theoretical (Rasch) function with an empirical function, which is created from the person's observed responses, evidence of misfit can be seen.

According to the Rasch model, the probability of a person correctly answering a dichotomously scored item (where 1 denotes a correct response and 0 denotes an incorrect response) is

$$\Pr(X_i = 1|\theta_n, \delta_i) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad (1)$$

In the model, θ_n represents the achievement level of person n and δ_i represents the difficulty level of item i.

Method

In operational testing situations, some person misfit is expected to exist, but the amount and type of person misfit is unknown. For the exploratory analysis planned in this study, the amount of misfit needed to be controlled, but the type of misfit did not. So, to produce a realistic simulated scenario, adaptive test data simulated to fit the Rasch model were generated. It was expected that some of the simulated person responses would misfit the model by chance. We classified each simulated person as fitting or misfitting the model using three person fit statistics: Outfit MSE (Wright and Stone 1979), Infit MSE (Wright and Masters 1982), and Between fit MSE, Bfit-P (Smith 1985). These will be described later. Then, we created and visually examined person response functions of two groups of test-takers: test-takers whose responses fit the model and test-takers whose responses did not fit the model.

Data Generation

To simulate an adaptive test, five hundred items were generated to represent a unidimensional item bank calibrated with the Rasch model using the catR package (Magis and Raiche 2012) for the R platform. These items were uniformly distributed over the logit range of −5.00 to 5.00. Next, using the same package (and the 500 items), dichotomous item responses were generated for 5000 test-takers drawn from a standard normal distribution, $\theta \sim N(0, 1)$. To simulate a dichotomous item response, a random number from a uniform distribution, $U(0, 1)$ was compared to the probability of giving the correct response computed from the known θ and δ, and the Rasch model (Eq. 1). When the probability of giving the correct response was greater or equal to the random number, the dichotomous response was set to 1; otherwise, it was set to 0 (Harwell et al. 1996).

Achievement level estimation was calculated with maximum likelihood procedures. A new provisional achievement estimate was computed after every response starting after three randomly selected items with approximate difficulties of −2.00, 0.00, and 2.00 were administered. The next item (i.e., item 4 and beyond) was selected based on the item's proximity to the current provisional estimate of achievement. This item selection process is the same as maximum information selection (Thissen and Mislevy 2000) when the Rasch model is used (Magis and Raiche 2012). No content coverage or item exposure constraints were placed on the item selection, and the test was stopped after 40 items were administered. To check that the maximum likelihood achievement estimator did not severely bias the results, we also performed the adaptive test process using the weighted maximum likelihood estimator (Warm 1989), which has been shown to correct for estimation bias when the number of items is small.

Data Analysis

The analyses for examining person fit were conducted using the final achievement estimates ($\hat{\theta}$) yielded from the adaptive test procedure, the known item difficulty values (δ) and the dichotomously scored item responses. First, the three person fit statistics designed to detect misfit to the Rasch model were computed: Outfit, Infit, and Bfit-P. The Outfit and Infit person fit statistics are useful for detecting random disturbances to the model, such as what may be produced by random guessing behavior or careless responding behavior (Smith and Plackner 2010). The Bfit-P person fit statistic is good for detecting more systematic disturbances, such as what may be produced by persons who run out of time, are sub-experts or have a deficiency in a content domain, or are slow to warm-up to the test (Smith and Plackner 2010). The formulations of these fit statistics are included below:

$$\text{Outfit } MSE_n = \frac{1}{L}\sum_{i=1}^{L}\frac{(X_{ni} - E_{ni})^2}{V_{ni}} \tag{2}$$

$$\text{Infit } MSE_n = \frac{\sum_{i=1}^{L}(X_{ni} - E_{ni})^2}{\sum_{i=1}^{L} V_{ni}} \tag{3}$$

$$\text{Bfit-P } MSE_n = \frac{1}{(J-1)}\sum_{j=1}^{J}\frac{\left(\sum_{i\in j}^{Lj} X_{ni} - \sum_{i\in j}^{Lj} E_{ni}\right)^2}{\sum_{i\in j}^{Lj} V_{ni}} \tag{4}$$

In these formulations, X_{ni} is the observed response for person n on item i (either 0 or 1), E_{ni} is the expected response for person n on item i based on the estimated achievement level (i.e., a probability calculated from the model), V_{ni} is the variance, i.e., $E_{ni}(1 - E_{ni})$, and L is the number of items on the test (Smith 1985). For the Bfit-P statistic, J is the number of item subsets and L_j is the number of items in each subset (Smith 1985).

The values for the three fit statistics can range from 0.00 to ∞ and are assumed to approximate a chi-squared distribution (Wright and Stone 1979; Smith 1991; Smith and Hedges 1982). The expected values of Outfit, Infit, and Bfit-P are 1.00 when the data fit the Rasch model. For Infit and Outfit, two types of response patterns are discordant with the Rasch model: muted and noisy response patterns. Fit statistics greater than 1.00 signify *noisy* fit. Noisy response patterns indicate that the response data are too unruly to be governed by the model. Substantively, this may indicate random responding or person dimensionality. Fit statistics less than 1.00 signify *muted* fit, which may suggest dependency in the data. Substantively, this may indicate item exposure (cheating) or very slow, methodical responding. In most person fit research, the concern is with identifying noisy response patterns (Reise and Due 1991), or those patterns with values substantially higher than 1.00. This concern was the focus of this study as well. Thus, the term *misfitting* referred to extreme person fit values located in the upper tail of the person fit statistic distribution.

In this study, the Bfit-P statistic was computed based on item administration order. In the context of this study, a large Bfit-P value would suggest that the person's performance on the first, middle, and/or last subsets of items do not accord with the model, rather than suggesting general misfit over the entire response pattern. The three subsets of items were as follows: items 4–15 (n = 12), items 16–27 (n = 12), and items 28–40 (n = 13). The first three items were omitted from the Bfit-P analysis because they were used to start the computer adaptive test.

Statistical Person Fit Analysis

Like in real testing situations, the responses that truly misfit the Rasch model in this study were unknown. A method to classify each person as fitting or misfitting the model was needed. This is done by identifying a critical person fit value and using it to classify persons as fitting or misfitting. One procedure that has been used and recommended in the recent literature uses a pre-set Type I error rate (i.e., α) and derives the critical value by simulating multiple sets of test data and taking the average fit statistic value at the pre-set point on the distribution (van Krimpen-Stoop and Meijer 2000; Lamprianou 2013; Petridou and Williams 2007). In this study, the process for establishing the threshold values for categorizing misfit followed a similar process to that reported in van Krimpen-Stoop and Meijer (2000), and used five replicated computer adaptive data sets with 10,000 persons each.

Graphical Person Fit Analysis

Fourteen persons who did not fit the Rasch model according to their fit statistics were chosen to illustrate person response functions. These persons were chosen because (a) they represented a range of estimated achievement levels, (b) they had only one of the three statistics flagged, and (c) they had a large fit statistic. The rationale for this last decision was because the response vectors in this study were simulated to fit the Rasch model, so by choosing the persons that produced large fit values, we were selecting those persons who were most likely to truly misfit. For comparison, 11 persons whose response patterns *fit* the Rasch model were also chosen. They were selected to represent a range of achievement levels.

Expected and empirical person response functions were created for these 25 test-takers. The expected PRF were created by plotting the expected probabilities, which were computed by inserting the final achievement estimate into Eq. 1 as θ_n and the item difficulty parameters as δ_i. The empirical PRF were created by smoothing the test-taker's original dichotomous responses, using an iterative Hanning procedure (Velleman and Hoaglin 1981). Specifically, the dichotomous responses to the items (y_i, first ordered by item difficulty) were transformed to continuous values between 0.00 and 1.00, s_i, by using the formula:

$$s_i = \left(y_{i-1} + 2y_i + y_{i+1}\right)/4. \tag{5}$$

In this formula, the first and last responses (i.e., y_1 and y_n) are left as-is. For the responses to items y_2 through y_{n-1} (items 2 through 39 in this study), s_i replaces the observed responses, y_i. The response y_i receives a weight of two and the two responses adjacent to y_i on each side receive a weight of 1.00.

The goal of the Hanning procedure was to obtain an adequate smooth, so that the final response function was useful, while still preserving the original response

pattern. To achieve this goal with dichotomous data, the procedure was repeated a number of times. Following Engelhard (2013), we iterated the procedure one time for each raw score point. That is, a person who obtained a raw score of 20 out of 40 had his or her responses smoothed 20 times.

The final smoothed values represented the *empirical* function, and were plotted on the same coordinate space as the expected function. Thus, for each person (N = 25), two PRF were created and superimposed on top of each other. The visual inspection of the PRF focused on the within-person congruence between the expected response function and the empirical response function.

Results

Before the person fit analyses commenced, the estimated achievement levels from the adaptive test procedure were evaluated for accuracy. Specifically, two sets of simulated results were evaluated: one that used maximum likelihood estimation and one that used weighted maximum likelihood estimation. The difference between the estimated and true achievement levels using these two methods were negligible, with the mean bias and mean absolute bias being discrepant at the thousandth decimal place or smaller. The achievement levels obtained from the maximum likelihood procedure were retained and used for the subsequent person fit analyses.

Statistical Person Fit Analysis

The threshold values for the fit statistics were established via simulation using a pre-set Type I error rate of 0.05. The resulting threshold values were Outfit = 1.166, Infit = 1.050, and Bfit-P = 2.997. Using these values, each test-taker was categorized as either fitting or misfitting the model three times, once for each person fit statistic. Person fit values greater than the threshold were defined as misfitting and person fit values less than the threshold were defined as fitting. The percentages of test-takers flagged for misfit for each fit statistic was 4.6 % for Outfit, 4.3 % for Infit, and 5.4 % for Bfit-P.

Graphical Person Fit Analysis

Person response functions for 14 misfitting test-takers and 11 fitting test-takers were evaluated. All 25 PRF were distinctive, however, similar characteristics were noticed. In the PRF, the x-axis represents the item difficulty continuum. The y-axis represents the probability of giving the correct response to the items, $[\Pr(x = 1)]$. The dichotomously scored items, i.e., the raw scored responses, are shown by the

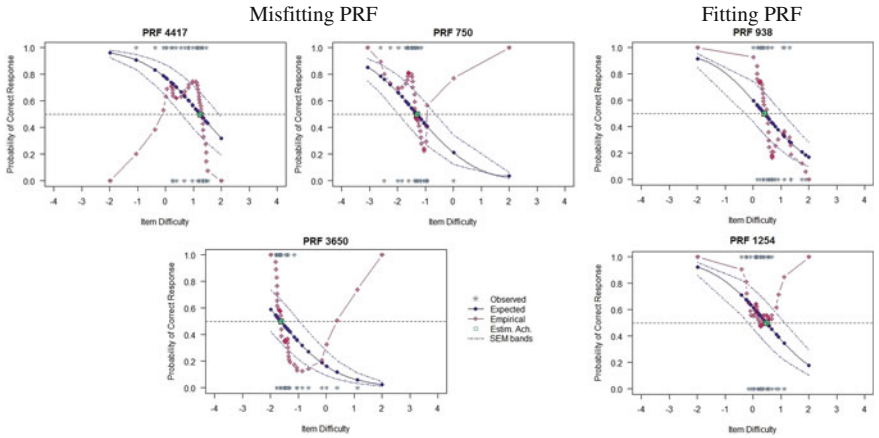


Fig. 1 Three person response functions illustrating misfit by Outfit. Two person response functions illustrating adequate fit are included for comparison. Note: The reference line included at Pr(x = 1) = 0.50 is where the item difficulty and person achievement level are equal. The *square* indicates the achievement estimate, $\hat{\theta}$, for these persons

asterisks, where a response of 1 means the test-taker gave the correct response to the item and a response of 0 means the test-taker gave the incorrect response to the item. The empirical response function created by the Hanning procedure in Eq. 5 is shown by the diamond line. The Rasch expected response function is shown by the circle line.

To aid in the interpretation of person misfit, three additional elements are included in the person response function plots. The estimated achievement level for the person is denoted by a square. The dotted lines on either side of the Rasch-expected function represent the SEM bands. These bands are calculated by plotting the Rasch probabilities for the estimated achievement level plus and minus two standard errors of measurement, i.e., $\Pr(x = 1|\hat{\theta} \pm 2 * \text{SEM})$. Finally, a horizontal reference line is drawn where the probability of the simulated person giving the correct response is 0.50, which is the location of the achievement estimate in the Rasch model.

Figure 1 highlights characteristics of misfit in the computer adaptive test as detected by Outfit. PRF 938 and 1254 illustrate adequate model fit and PRF 4417, 3650, and 750 illustrate Outfit misfit. One common observation for the persons flagged as misfitting in Fig. 1 is the large unexpected correct (or incorrect) response at the end (or beginning) of the response pattern. For instance, for Person 4417, the diamond point located at item difficulty −2.00 illustrates that this person gave an incorrect answer to this item and the circle point at the same location illustrates that the model expected the person to give a correct answer.

A second common observation for the persons flagged as misfitting by Outfit is the unexpected responses in the middle of the functions. There are some peaks and dips in the empirical response function located in the middle of the item difficulty

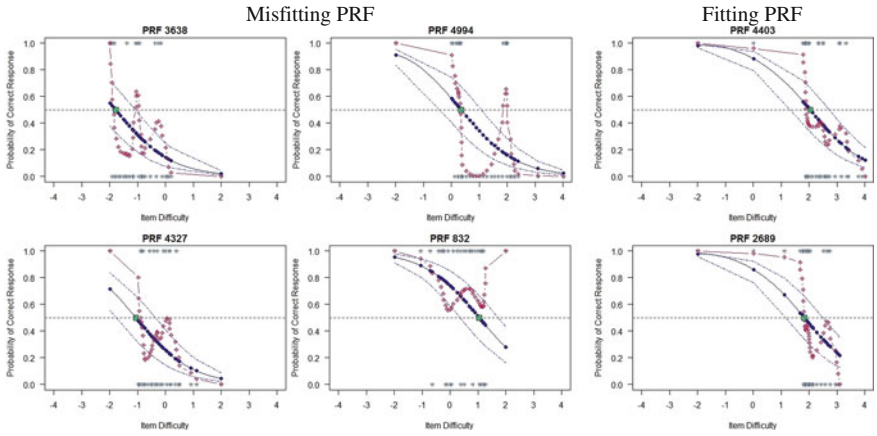


Fig. 2 Four person response functions illustrating misfit by Infit. Two person response functions illustrating adequate fit are included for comparison. Note: The reference line included at Pr (x = 1) = 0.50 is where the item difficulty and person achievement level are equal. The *square* indicates the achievement estimate, $\hat{\theta}$, for these persons

continuum, near the estimated achievement level, and a segment of the empirical function extends beyond the SEM bands. These visual characteristics of the PRF are consistent with judgments of person misfit, as suggested by the person fit statistics.

It is noted that the empirical function for Person 938, who was categorized as fitting the model, also extends beyond the SEM bands around the achievement estimate, and the empirical function for Person 1254 exhibits an unexpected response at the end of the responses. The difference is that for these fitting persons, *either* peaks and dips *or* an extreme unexpected response is present in the empirical function, but not both. For the misfitting persons, severe dips and peaks of the empirical function are seen *in addition to* an unexpected response at the end (or beginning).

Figure 2 highlights characteristics of misfit in the computer adaptive test as detected by Infit. PRF 4403 and 2689 illustrate fit to the model. PRF 3638, 4327, 4994, and 832 illustrate misfit to the model. Here, the common observation for the persons flagged as misfitting is the extreme discrepancies between the expected and empirical functions around the probability of 0.50 or the estimated achievement level. For instance, large dips and peaks are displayed for Persons 3638, 4327, and 4994 instead of a steady decline in the probabilities. A substantial portion of the empirical functions for these misfitting persons extend beyond the SEM bands. By contrast, the empirical function for Persons 4403 and 2689 (who fit the model) exhibit smaller dips and peaks, which barely fall beyond the SEM bands. For Person 832, the empirical function does not approach the probability of 0.50. These visual characteristics of the PRF are consistent with judgments of person misfit and suggest that more than one achievement estimate is plausible.

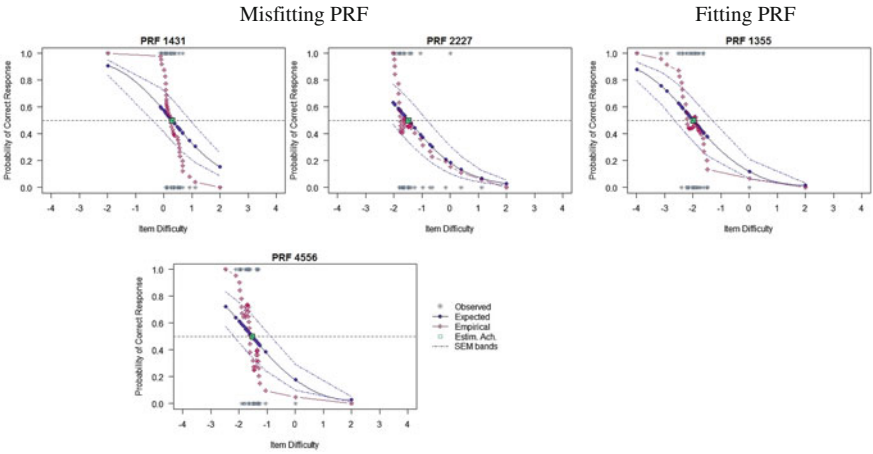


Fig. 3 Three person response functions illustrating misfit by Bfit-P. One person response function illustrating adequate fit is included for comparison. Note: The reference line included at Pr (x = 1) = 0.50 is where the item difficulty and person achievement level are equal. The *square* indicates the achievement estimate, $\hat{\theta}$, for these persons

Figure 3 highlights characteristics of misfit in the computer adaptive test as detected by Bfit-P. Person response function 1355 illustrates fit to the model and PRF 1431, 4556, and 2227 illustrate misfit to the model. Here, there is no obvious or consistent characteristic that shows misfit. For Persons 1431 and 4556, the empirical functions appear steep, but they do not show major unexpected deviations (i.e., dips and peaks) from the expected PRF. For Person 2227, the empirical function follows the expected function well. The characteristics of these PRF are not consistent with judgments of person misfit, as suggested by the person fit statistics.

Discussion and Conclusion

In this study, the simulated persons fit the model generally well, but there was some evidence of individual person misfit. This observation highlights the need for conducting individual person fit analyses in practice. Although achievement estimates from IRT models have been shown to be fairly robust to model-data misfit in paper-pencil tests (Adams and Wright 1994; Sinharay and Haberman 2014) and CAT (Glas et al. 1998), test-takers in real situations may respond to items in unique and unstudied ways that may threaten the inferences of their scores. Moreover, in CAT where the item parameters are considered known, and where each test-taker may receive a different set of items, evaluating individual person fit can provide

vital information about model-data fit that may be absent from pre-test and post-test quality checks.

The advantage of using a two-step, statistical and graphical, procedure for examining individual person fit in CAT is that it allows for a statistical quantification about individual person fit (i.e., person fit statistic) to be further informed by a visual inspection of the actual response pattern (e.g., empirical person response function). Given the skepticism regarding the use of existing person fit statistics in an adaptive test context (Glas et al. 1998; McLeod and Lewis 1999; Nering 1997; van Krimpen-Stoop and Meijer 1999), this additional information is warranted.

Because they enhance validity evidence for score interpretation and use (APA/AERA/NCME 2014), individual person fit techniques can improve computer adaptive testing practice. Person response functions require experience and some subjective judgement to interpret, but they provide complementary information about person fit. These visual representations of misfitting patterns may help researchers and other educational stakeholders understand the substantive implications of person misfit. For instance, in this study, the PRF showed *why* misfit (with the Outfit and Infit statistics) was detected and the general location. For the Bfit-P statistic, the PRF showed no substantive misfit, which suggests that more information about these person responses should be gathered before a judgement about person fit is made. Practitioners and researchers can use these two pieces of information to evaluate potential threats to a person's test score inference.

References

Adams, R. J., & Wright, B. D. (1994). When does misfit make a difference? In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 244–270). Norwood, NJ: Ablex.

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

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*(443), 910–919.

Chang, H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics, 37*, 1466–1488.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10*(1), 101–119.

Engelhard, G, Jr. (2013). Hanning (Smoothing) of person response functions. *Rasch Measurement Transactions, 26*(4), 1392–1393.

Ferrando, P. J. (2014). A general approach for assessing person fit and person reliability in typical-response measurement. *Applied Psychological Measurement, 38*, 166–183.

Glas, C. A., Meijer, R. R., & van Krimpen-Stoop, E. M. (1998). *Statistical tests for person misfit in computer adaptive testing (Research Report 98–01)*. The Netherlands: University of Twente, Faculty of Educational Science and Technology.

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101–125. doi:10.1177/014662169602000201.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person fit statistics. *Applied Measurement in Education, 16*, 227–298.

Lamprianou, I. (2013). The tendency of individuals to respond to high-stakes tests in idiosyncratic ways. *Journal of Applied Measurement, 14*(3), 299–317.

Magis, D., & Raiche, G. (2012). Random generation of response patterns under computer adaptive testing with the R Package catR. *Journal of Statistical Software, 48*(8), 1–31.

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 43*, 147–160.

Meijer, R. R. (2005). *Robustness of person-fit decisions in computerized adaptive testing (Computerized Testing Report 04–06)*. Newtown, PA: Law School Admission Council.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.

Meijer, R. R., & van Krimpen-Stoop, E. M. (2010). Detecting person misfit in adaptive testing. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 315–329). New York, NY: Springer.

Nering, M. L. (1997). Distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 127–155.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement, 22*, 53–69.

Perkins, A., Quaynor, L., & Engelhard, G. (2011). The influences of home language, gender, and social class on mathematics literacy in France, Germany, Hong Kong, and the United States. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 4*, 35–58.

Petridou, A., & Williams, J. (2007). Accounting for aberrant test response patterns using multilevel models. *Journal of Educational Measurement, 44*(3), 227–247.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research, Expanded edition, Chicago: University of Chicago Press (Original work published 1960).

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*, 217–226.

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice, 33*, 23–35.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45*, 433–444.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51*, 541–565.

Smith, R. M., & Hedges, L. V. (1982). Comparison of likelihood ratio $\chi^2$and Pearsonian χ^2 tests of fit in the Rasch model. *Education Research and Perspectives, 9*, 44–54.

Smith, R. M., & Plackner, C. (2010). The family approach to assessing fit in Rasch measurement. In M. Garner, G. Engelhard Jr., W. Fisher Jr., & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 1, pp. 64–85). Maple Grove, MN: JAM Press.

Trabin, T. E., & Weiss, D. J. (1979). *The person response curve: Fit of individuals to item characteristic curve models*. (Research Report 79–7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 101–133). Hillsdale, NJ: Erlbaum.

van Krimpen-Stoop, E. M., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327–345.

van Krimpen-Stoop, E. M., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201–219). Dordrecht, The Netherlands: Kluwer Academic Publishers.

van Krimpen-Stoop, E. M., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199–217.

Velleman, P. F., & Hoaglin, D. C. (1981). Smoothing data. In P. Velleman & D. Hoaglin (Eds.), *Applications, basics, and computing of exploratory data analysis* (pp. 159–199). Boston, MA: Duxbury Press.

Walker, A. A., Engelhard, G. Jr., Royal, K. D., & Hedgpeth, M. W. (2016). Exploring aberrant responses using person fit and person response functions. *Journal of Applied Measurement*, *17*(2), 194–208.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

The Influence of Repetition Type on Question Difficulty

Paul Horness

Introduction

There is a plethora of factors that influence listening comprehension such as speaking speed, accent, or topic knowledge. One aspect that has not been examined fully is repetition type. When investigated, the effects of repetition in testing listening comprehension have been mixed for several reasons. The primary reason is that repetition has been defined and operationalized too vaguely. Some researchers have defined it as just repeating the stimuli. Another reason is that repetition has also not been the primary focus of a study; rather it has been one of the several research questions investigated.

The theory of the spacing effect (Ebbinghaus 1885/1913) is the finding that memory performance is better when repetitions are separated by other items (i.e., spaced repetition), than when the repetitions immediately follow one another (i.e., massed repetition). The *lag effect* is the amount of time between the repetition of items; research has indicated that a longer lag between repetitions increases recall. Often the spacing effect and lag effect are confused with each other. The spacing effect refers to the comparison between a massed and spaced method, whereas the lag effect refers to the different retrieval times used in the spaced method.

The use of repetition is common in our daily lives. The spacing effect commonly used in commercials, advertising, speeches, or public announcements. In teaching, the spacing effect is often associated with the long-term effects of repetition, namely recycling material over a course or material in spiraled curriculum. The spacing effect, however, has benefits in the short term as well as the long term. Greater awareness of the spacing effect would enhance its application in education.

P. Horness (✉)
Atomi University, Niiza, Japan
e-mail: tokyopaul93@me.com

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_25

L2 Memory Processes

Second language listening processes do not differ from first language listening processes in any physical aspect (Kroll et al. 2012) except that processing capacity is generally reduced (Call 1985; Faerch and Kasper 1986; Hasegawa et al. 2002). The difficulty for second language learners arises in comprehending specific elements of the language, and any necessary compensation, such as using background knowledge to modify the deficiencies, provide another opportunity for miscomprehension. Even two native speakers encounter misunderstandings and do not accurately comprehend everything they hear at all times. Compensatory skills, such as using visual cues or world knowledge, can help listeners compensate for incomplete listening comprehension.

Mackey et al. (2010) found that working memory accounted for less than 20 % of the variance in measuring oral output, a finding that indicates that other factors contribute to the output. Sunderman and Kroll (2009) suggested that working memory for language learners, who are beyond the basic elements of the language, is the ability to control attention and suppress competing processes. The ability to control attention explains why high L1 working memory capacity does not transfer directly to high L2 working memory capacity; the L1 might interfere with the L2. As numerous studies indicate (De Bot and Kroll 2010; Costa 2005; Dijkstra 2005; Marian and Spivey 2003; Schwartz et al. 2007), L2 learners cannot suppress L1 activations even when they are not in use. This finding holds for reading, listening, and planned speaking. These studies indicate that the parallel activation of both languages influence working memory capacity.

Spacing Effect in L2

Serrano and Muñoz (2007) outlined several studies where intensive learning was more productive than spaced learning. In vocabulary acquisition, Collins et al. (1999) found that vocabulary items learned over five months were better recalled that items learned over after 10 months. In speaking, several studies (Freed et al. 2004; White and Turner 2005) indicated that fluency increased over a short intensive time than over a long period. In listening, several studies indicated that intensive or compressed classes helped listening comprehension (Lapkin et al. 1998; Lightbown and Spada 1994). Although these studies demonstrate intensive study can be more effective than traditional instruction, there is a matter of how the terms *spaced* and *massed* were defined. Most of the studies were based on learning over a year. For example, one class had instruction for 350 h over 10 months versus 350 h over 5 months. Or, the amount of instruction time differed between the conditions. For example, one intensive course had 350–400 h of instruction over the year with 18–20 h per week, but the traditional course only had 120 h of instruction over the year with a maximum of 4 h a week. These results may not reflect the underlying working of the spacing effect due to the lag effect.

There have been several studies focusing on L2 listening comprehension using repetition. Brindley and Slatyer (2002) explored the influence of tasks on test results involving the Certificates in Spoken and Written English (CSWE). The principal aim was to identify how five key task characteristics and task conditions affected test difficulty. One of the variables examined was the number of hearings. The participants were 284 adult ESL learners enrolled in Certificate III in teaching centers across three states in Australia. These participants engaged in various combinations of three tasks. One task (control condition) was given to all the participants. The participants listened once to a two-minute recorded monolog concerning the Australian educational system, and then completed a sentence-completion task in which a few words of English were written in ten sentences. Four versions of the remaining two tasks were created based on the item difficulty variable; these tasks were randomly assigned after the participants had completed the first task. Other than the topic change, the second task's baseline was similar to the first task, but the other versions included changing the item format, repetition, and the use of live speech. The third task's baseline was similar to the control version, but the topic was about the Guide Dog Association in Australia. Additionally, the third task's versions included short answers instead of sentence completion, a dialog instead of a monolog, and a faster speech rate.

In order to determine which tasks influenced item difficulty, the researchers first analyzed the scores using a Rasch-based program to obtain person ability and item difficulty estimates. In order to do this, all of the test forms were combined and treated as a single test containing 89 items with task 1 (control) providing a linking set of common items. Based on the ability/difficulty scale, the tasks could be interpreted as easier and more difficult. The easiest task required the participants to complete a table after listening. The most difficult task was characterized by increased speech rate. Although not addressed specifically, the results from a graph indicated that the number of hearings did not affect item difficulty when compared with other stimuli variables such as listening once, or live versus recorded speech.

Chang and Read (2006) investigated four listening support formats in which one of the conditions was repeated input. The first research question concerned whether different types of listening support would affect listening performance. The second research question asked whether the listening support types would affect higher or lower proficiency participants in the same manner. They examined the effects of the different formats on listening comprehension with 160 students from intact classes studying business at a college in Taipei, Taiwan. The participants were given one test condition based on their class and class day. Based on a TOEIC test, each group was further divided into low and high listening proficiency sub-groups. The participants in each condition completed two listening tests with 15 multiple-choice questions for each listening text. In the repeated input condition, the students were asked to listen to the text without any special preparation. Then they previewed questions before listening to the text twice, so they heard the text three times in all. Thereafter they answered 15 multiple-choice questions in three minutes. The steps were repeated for the second listening text.

Chang and Read (2006) conducted a 4 × 2 ANOVA. The dependent variable was the combined test score. The independent variables were four types of listening support (previewing questions, repeated input, topic preparation, and vocabulary instruction) and two listening proficiency levels (high and low). The results indicated that repeated input generated the second highest mean test scores while topic preparation was the highest. For the first research question, significant main effects were found for listening support and listening proficiency. There was also a statistically significant interaction between listening support and listening proficiency. Comparing the two proficiency groups, two of the four types of listening support, previewing questions and repeated input were statistically significant. The high proficiency group scored higher using these two types of listening support than the lower proficiency group. Their results indicated that the different types of listening support affected comprehension scores for the different proficiency groups, but the effect sizes were small.

In answer to the second research question, the researchers reported that listening support activities affected the low and high proficiency levels differently. Their results indicated that high proficiency learners benefitted the most from repeated input, but the differences between two of the three listening support activities were not statistically significant. The low-proficiency learners benefitted the most from topic preparation, but the difference of this activity from repeated input was not statistically significant whereas the other listening support activities were.

Chang and Read (2007) examined listening support factors on listening comprehension with 140 students at a five-year post secondary educational program at a Taiwanese college with low levels of listening proficiency as measured by the TOEIC listening section (a scaled score of 165 out of 465). Two research questions were investigated. The first asked was what type of listening support, repetition, visual, or textual, would enhance comprehension for low-proficiency listeners. The second asked what type of support (visual, textual, or repeated input) would affect the students' perceptions of the listening task. Chang and Read stated that learners bring beliefs to the task, and those beliefs influence task performance.

They conducted a one-way ANOVA. The independent variables were visual support, textual support, and repeated support. The dependent variable was the listening comprehension score. The results indicated that all three types of listening support resulted in significantly higher scores than the control condition, and that repeated input produced significantly higher scores than the other types of listening support.

Cognitive Difficulty of Questions

One aspect that influences cognitive item difficulty is the interaction between the item type and test-taker. For instance, the test item and its relationship to the participant's background knowledge plays a role in determining question difficulty. Yi'an (1998) concluded that background knowledge played a role in how the

participants answered the questions. For higher proficiency learners answering multiple-choice questions, background knowledge acted as a facilitator as it allowed the learners to use the stem questions or distractors when responding. However, as Yi'an pointed out, this facilitating effect does not guarantee a correct response. For lower proficiency learners, multiple-choice questions acted in a compensatory way to fill in missing information. Again, this effect did not necessarily mean that the learners responded correctly.

Another aspect is the relationship of the question stem and response item. Nissan et al. (1996) concluded that inference items, which ask about information not stated explicitly in the passage, were significantly more difficult than items that required information explicitly stated in the passage. Kostin (2004) replicated Nissan et al's. (1996) study and came to the same conclusion. In addition, both studies indicated that lexical overlap and redundancy helped comprehension. Further studies have confirmed these findings (Brindley and Slatyer 2002; Buck and Tatsuoka 1998).

For this study, question difficulty was considered in the following ways. First, Bloom's taxonomy (1956) of six levels of cognitive difficulty was used as the basis for determining question difficulty. Brown (2001) interpreted Bloom's taxonomy for language purposes and outlined seven levels (p. 172). The first level, which is considered the easiest level, is called knowledge questions. These types of questions ask for factual information, and test recall and recognition of information. The second level, which is considered more cognitively difficult, is comprehension questions. These types of questions ask individuals to interpret and infer information. The fourth level, which is more difficult than the preceding levels, is called inference questions. These questions include forming conclusions that are not directly stated in the input. Second, Henning's (1991) definition was also adapted to make comparison more applicable. Therefore, lower order cognitive difficulty was defined as questions requiring an understanding of specific information stated in the passage within a single sentence. Higher order cognitive difficulty was defined as questions that can only be answered by using information from two or more sentences or inferences.

Methods

The participants were first and second year students attending a Japanese national university. The students are streamed into their course based on the institution's TOEIC test. All of the participants' TOEIC listening scores were under 300. The 242 participants were from six intact classes taking a required TOEIC course. One class from the first year and one class from the second year was randomly grouped together and given one of three treatment conditions. Table 1 gives the descriptive statistics. Although the group's scores were not equal and a one-way ANOVA indicated a statistical difference between the massed condition and the other conditions, it does not necessarily indicate that the participants were at different proficiency levels for several reasons. First, the TOEIC standard error of measurement

Table 1 TOEIC scores for each group

Condition	Control	Massed	Spaced
N	71	72	99
M	168.73	196.46	174.09
SD	25.74	33.39	29.75

is ±25 points (TOEIC 2016a), so the control group and spaced group can be considered equal. Second, all of the TOEIC listening scores were below 300 and the average below 200, so based on TOEIC's listening descriptions (TOEIC 2016b), the participants' listening proficiency were roughly equal. Third, the university divides the classes based on the complete TOEIC score. As TOEIC (2016a) indicates, there is a high correlation between the scores. For example, the listening score correlates at approximately 0.8 to the reading score while listening score and the total score correlates at 0.9.

Toward the end of the semester in order to prepare for the semester-ending TOEIC test, the students took the teacher-made listening test over three weeks. Ten listening passages with five multiple-choice questions for each passage were created. Each passage is approximately one minute in length. There were five passages each of monologs and dialogs. The five multiple questions were created with two distinct types of questions, specific detail and inference. A key difference between the TOEIC test and the teacher-made test was the use of distractors. Commonly, TOEIC uses terms or phrases from the listening text as distractors. For this test, none of the distractors were from the listening text. The scores were not counted as part of the grade.

The procedures for each listening passage are shown in Table 2. The topic is introduced prior to listening to induce schema building in all the listening conditions. In the control condition, students listen to the passage once and then answer five questions. In the massed repetition condition, students listen to the passage twice and then answer five questions. In the spaced repetition condition, students listen to the passage, count 5–1 to interrupt the phonological loop, which will interrupt their working memory, listen a second time, and then answer five questions. The students were encouraged to take notes while listening, but were not allowed to look at the questions while listening.

Table 2 Procedures for each condition

Condition	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
Control	Show topic	Play passage	Show questions	Answer questions		
Massed repetition	Show topic	Play passage	Play passage again	Show questions	Answer questions	
Spaced repetition	Show topic	Play passage	Students count 5–1 (intervening variable)	Play passage again	Show questions	Answer questions

Results

All Rasch analyses were conducted with WINSTEPS version 3.92.

The Rasch model allows us to examine all the items on one form of measurement, i.e., a ruler. For this study, all of the items under all three conditions were combined and analyzed under one measurement. Therefore, if the hypotheses hold true, the Rasch Model should indicate bias based on the condition the item was under. The raw mean score for the participants was 47.5 with a standard deviation of 1.6. The Rasch mean estimate was 50.0 CHIPS, item reliability estimate was 0.98, and item separation was 6.65. The Rasch person reliability estimate was 0.81, and person separation was 2.08. The person separation number is a little surprising since the participants were filtered into groups based on the TOEIC examination. Further examination indicated that the participants were not from one treatment condition.

Rasch Fit Statistics

The criteria for the fit statistics were set at ±0.7–1.5 for the mean squares. Checking outfit scores first, all fifty items were within the set criteria except for item 21 and item 4. Those items mean squares were 1.70 and 1.63, respectively, with standard z-scores of 1.5 and 1.9, respectively. Checking infit scores next, all items were within the set criteria.

Figure 1 shows the Wright map, which is the person–item relationship in a pictorial representation (Bond and Fox 2007). The CHIPS scale is shown on the far left side of the figure. According to Linacre (2008) CHIPS are a useful transformation, in which 1 logit = 4.55 CHIPS. In this user-scaling system, standard errors tend to be about 1 CHIP in size. A comparison of the locations of the person measures (left side) and item measures (right side) shows that the mean CHIPS for the person measures for the participants (M = 47.50; SD = 1.6) is equal to the mean of the item measures (M = 50.00; SD = 0.80). In addition, Fig. 1 shows several of the item measures were redundant in that they shared the same location on the logit scale with at least one other item. Nonetheless, the item measures were spread out sufficiently in that the items range were beyond the participants' listening comprehension ability, except for only a few participants, on this test. Along the left side of the map, the participants are spread out over 22 CHIPS, minimum = 33.5 CHIPS, maximum = 55.8 CHIPS, with higher proficiency students toward the top of the map and lower proficiency students toward the bottom. Along the right side of the map, the items spread out over 23 CHIPS, minimum = 40.3 CHIPS, maximum = 63.5 CHIPS, with the easier items toward the bottom of the map and the more difficult items toward the top. In this case, the listening comprehension test covers the abilities of the highest students, but there are a few low-proficiency students that the test did not cover. The common linear interval data for persons and

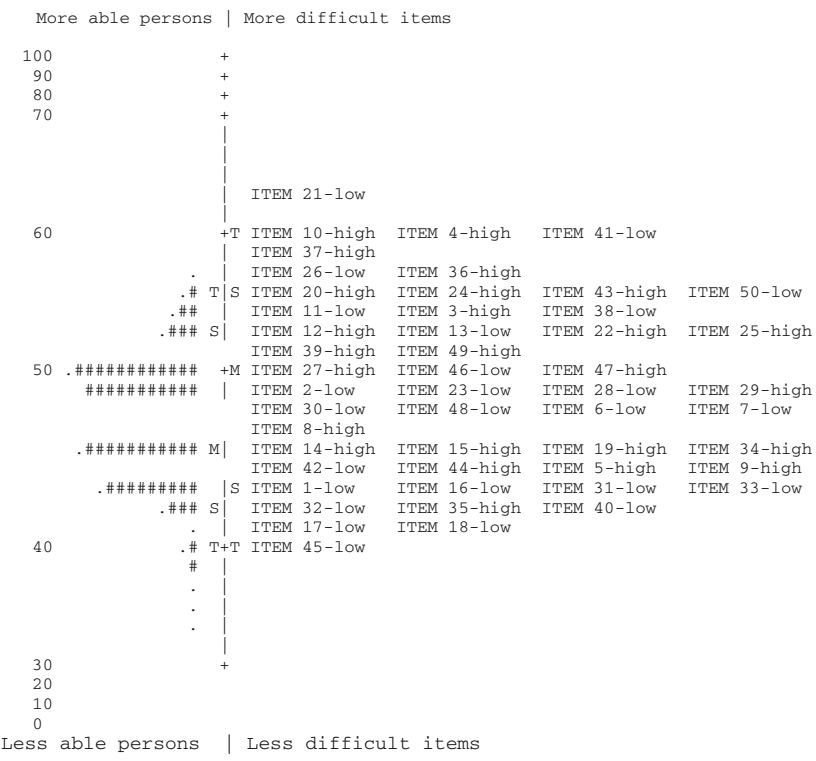


Fig. 1 Wright map of items. *Note* Each "#" is 4 participants; Each "." is 1–3 participants

items gives a clear demonstration of whether the items matched the persons' abilities for the construct measured. In this population sample, the mean of the items were slightly above the participants' ability.

The Rasch principal components analysis of the residuals was applied to detect other potential measurement dimensions in the listening test data. The Rasch model accounted for 25.2 % of the variance in the data. The first residual contrast, Fig. 2, had an eigenvalue of 4.0. Upon further examination, five items separated from the other items with loadings over +0.5. All five items were from the last listening passage. Although the test was given over three periods to reduce fatigue and help concentration, the last passage stands out against the other passages.

The second residual contrast had an eigenvalue of 3.2. There was no clear break between the items, and most of the loadings were under 0.5. The top three items were questions of high difficulty while the bottom three questions were low difficulty.

A Differential Item Function (DIF) analysis investigates the different characteristic of a test item between subpopulations and is useful in finding biased items toward a particular subpopulation. In this study, it was hypothesized that spaced repetition would increase comprehension scores greater than the control group and

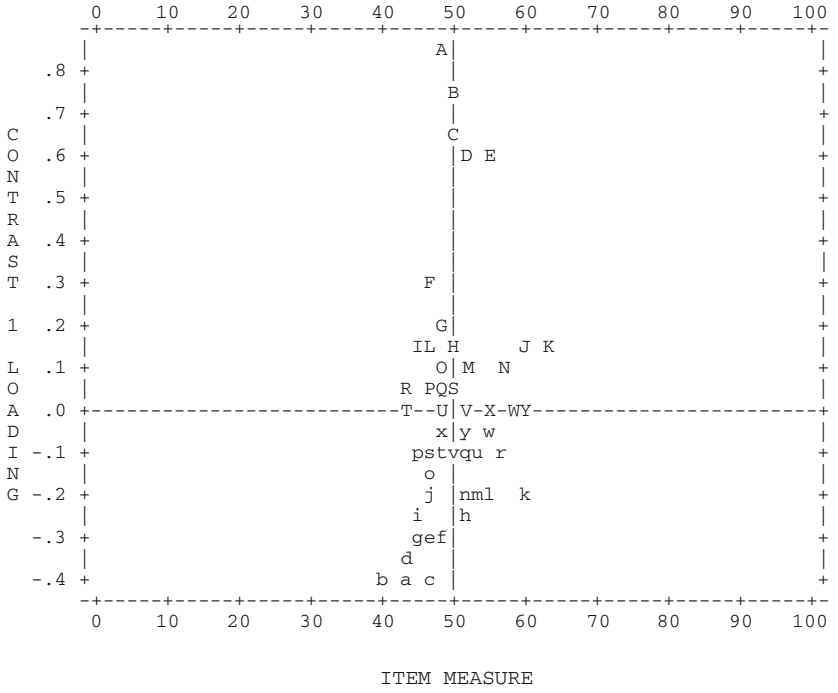


Fig. 2 First contrast loading

the massed repetition group. Therefore, bias should exist in the model, favoring spaced repetition the most. There are two parts that need to be examined when analyzing DIF. The first is the logit needs to be big enough to examine and therefore logits greater than 0.5 are examined. Second, the Rasch-Welch *t*-value needs to be greater than 2.0 as to indicate that the significance did not happen by chance. Overall, as indicated in Fig. 3, there appears to be bias for the spaced repetition group. However, the bias is not consistently favorable, nor is it significant. For example, with the first item, the spaced repetition group found it much easier to comprehend than the other groups. The DIF contrast for item 1 is 4.0 between the control group and the spaced repetition group, while the massed repetition group is 0.2 between the control group. Additionally, the *t*-value is 2.60 for the spaced repetition group. However, with item 15, the spaced repetition group found it more difficult than the other groups. The DIF contrast for item 15 is −3.4 between the spaced repetition group and the control group with a *t*-value of 2.17, while the massed repetition group is 0.1. By examining each item using these above guidelines, only items 1 and 15 were biased. The remaining items might have logits greater than 0.5, but the *t*-values were less than 2.0. Overall, the items in the spaced repetition condition were inconsistent. About 22 items were easier in the spaced repetition condition than the other conditions while about 22 were more

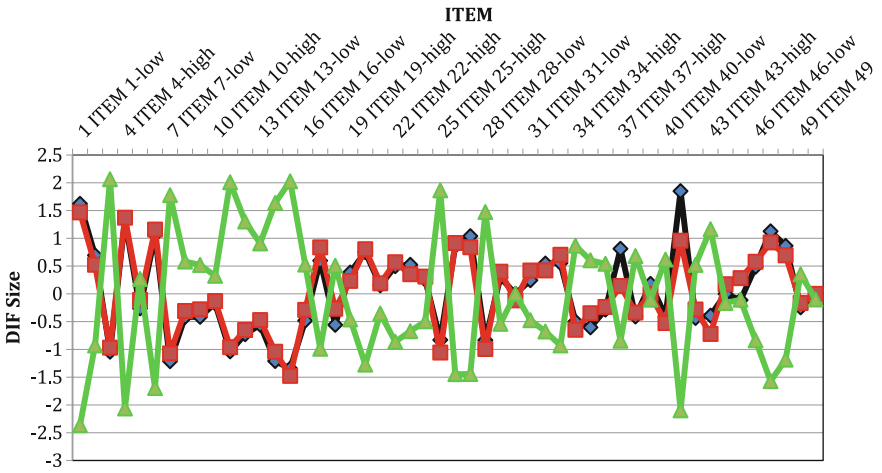


Fig. 3 Measure for the differential item function. *Note Black line* Control group; *Red line* Massed repetition group; *Green line* Spaced repetition group; *low* Low cognitive difficulty; *high* High cognitive difficulty

difficult. Additionally, the difficulty of the question was not favored for any condition either. For the most part, the control group and massed repetition groups mirrored one another and did not have any significant bias.

Discussion

The purpose of this study was to investigate the effect of repetition on question difficulty on a listening comprehension test. Two types of repetition, spaced and massed, were given on a listening comprehension test in order to examine the affect on question difficulty. As with previous research (Brindley and Slatyer 2002), massed repetition had a similar effect as the control group, i.e., listening once. It was hypothesized that the spacing effect would have a greater positive effect, but overall, the results did not indicate greater listening comprehension. Nor did the spacing effect affect question difficulty in any significant way. The spacing effect had an effect on question difficulty, but it is not clear as to why it was positive for some items and negative for others.

References

Bloom, B. (1956). *Taxonomy of educational objectives*, handbook I: The cognitive domain. New York: David McKay Co. Inc.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd). Mahwah, NJ: Erlbaum.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*, 369–394.

Brown, H. D. (2001). *Teaching by principles: An interactive approach to language pedagogy*. White Plains, NY: Longman.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119–157.

Call, M. E. (1985). Auditory short-term memory, listening comprehension, and the input hypothesis. *TESOL Quarterly, 19*, 765–781.

Chang, A., & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly, 40*, 375–397.

Chang, A., & Read, J. (2007). Support for foreign language listeners: Its effectiveness and limitations. *RELC Journal, 38*, 375–395.

Collins, L., Halter, R. H., Lightbown, P. M., & Spada, N. (1999). Time and the distribution of time in L2 instruction. *TESOL Quarterly, 33*(4), 655–680.

Costa, A. (2005). Lexical access in bilingual production. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 308–325). New York: Oxford University Press.

De Bot, K., & Kroll, J. F. (2010). Psycholinguistics. In N. Schmitt (Ed.), *An introduction to applied linguistics* (2nd ed., pp. 124–142). London: Hodder Education.

Dijkstra, T. (2005). Bilingual word recognition and lexical access. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 179–201). New York: Oxford University Press.

Faerch, C., & Kasper, G. (1986). The role of comprehension in second language learning. *Applied Linguistics, 7*, 257–274.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *SSLA, 26*, 275–301.

Hasegawa, M., Carpenter, P. A., & Just, M. A. (2002). An fMRI study of bilingual sentence comprehension and workload. *Neuroimage, 15*, 647–660.

Henning, G. (1991). A study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on TOEFL listening comprehension item performance. Retrieved from NJ: Educational Testing Service. Website: http://www.ets.org/toefl/research/archives/research_report.

Kostin, F. (2004). Exploring item characteristics that are related to the difficulty of TOEFL dialogue items. *Research Reports 79*. Educational Testing Service, Princeton, New Jersey.

Kroll, J. F., Bogulski, C. A., & McClain, R. (2012). Psycholinguistic perspectives on second language learning and bilingualism: The course and consequence of cross-language competition. *Linguistic Approaches to Bilingualism*, *2*, 1–24.

Lapkin, S., Hart, D., & Harley, B. (1998). Case study of compact core French models: Attitudes and achievement. In S. Lapkin (Ed.), *French second language education in Canada: Empirical studies* (pp. 3–31). Toronto, ON: University of Toronto Press.

Lightbown, P. M., & Spada, N. (1994). An innovative program for primary ESL students in Quebec. *TESOL Quarterly, 28*(3), 563–579.

Linacre, J. M. (2008) A user's guide to WINSTEPS: Rasch measurement computer program. Chicago: Mesa Press.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning, 60*, 501–533.

Marian, V., & Spivey, M. J. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition*, *6*, 97–115.

Nissan, S., DeVincenzi, F., & Tang, L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension (TOEFL Research Report RR 95-37)*. Princeton, NJ: Educational Testing Service.

Schwartz, A. I., Kroll, J. F., & Diaz, M. (2007). Reading words in Spanish and English: Mapping orthography to phonology in two languages. *Language and Cognitive Processes, 22*, 106–129.
Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System, 35*, 305–321.
Sunderman, G., & Kroll, J. (2009). When study abroad fails to deliver: The internal resources threshold effect. *Applied Psycholinguistics, 30*, 79–99.
TOEIC. (2016a). TOEIC's user guide. Retrieved from https://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf
TOEIC. (2016b). Listening score description. Retrieved from https://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_Score_Desc.pdf
White, J., & Turner, C. E. (2005). Comparing children's oral ability in two ESL programs. *Canadian Modern Language Review, 61*(4), 491–517.
Yi'an, W. (1998). What do tests of listening comprehension test?—A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*, 21–44.

A Comparison of Methods for Dimensionality Assessment of Categorical Item Responses

Chen-Wen Liu and Wen-Chung Wang

Exploratory factor analysis (EFA) and principal component analysis (PCA) are two common methods for dimensionality assessment. Both methods assume that the variables are continuous. In EFA, a few common factors are extracted from the variables; whereas in PCA, a few components are created to account for the variations among variables. Linear EFA or PCA was enhanced by means of parallel analysis (PA). In the PA method, a large number (e.g., 100) of datasets that have the same size of response matrix of the real dataset are randomly simulated and then analyzed with linear EFA or PCA. These eigenvalues obtained from random datasets are compared with those of the real dataset. The number of factors is the number of times when the eigenvalues derived from the real dataset are larger than the 95th percentile (or mean) of the eigenvalue distribution of the simulated datasets. The PA method appears promising in dimensionality assessment for continuous, dichotomous, and polytomous variables, and it can accommodate the Pearson correlation and the polychoric correlation (Cho et al. 2009; Tran and Formann 2009; Weng and Cheng 2005).

Weng and Cheng (2005) concluded that the PA method performs well for unidimensional dichotomous variables when the 95th or 99th percentile of the random data eigenvalues criteria is used. Cho et al. (2009) observed that the PA method based on the Pearson correlation performs at least as well as that based on the polychoric correlation in most conditions. However, Tran and Formann (2009) showed that the PA method does not perform well in assessing the dimensionality of dichotomous variables based on the Pearson correlation or the tetrachoric correlation. Finch and Monahan (2008) implemented a modified PA method based on nonlinear factor analysis with the TESTFACT program (Wilson et al. 1991) and

C.-W. Liu (✉) · W.-C. Wang
Education University of Hong Kong, Tai Po, Hong Kong
e-mail: cwliu@eduhk.hk

W.-C. Wang
e-mail: wcwang@eduhk.hk

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_26

parametric bootstrap sampling, and they concluded that the new method outperforms the DIMTEST method in identifying the unidimensional structure. Timmerman and Lorenzo-Seva (2011) implemented the PA method with a minimum rank factor analysis to assess the dimensionality of polytomous variables, and they observed that this method outperforms traditional PA methods in most conditions. They thus recommended the use of the polychoric correlation for polytomous variables if there is no convergence problem. In this study, the PA method implemented in the computer program FACTOR (Lorenzo-Seva and Ferrando 2006) was used to serve as a baseline, with which other methods were compared.

The Hull method can be regarded as a generalization of the scree test (Lorenzo-Seva et al. 2011). A series of EFAs are conducted, each with a different number of factors, starting from 1 to an upper bound. The upper bound can be determined by the indicated number of factors via the PA method plus one for the search range of the elbow on the scree plot. Each run has several goodness-of-fit indices (e.g., the comparative fit index, root mean square error of approximation, standardized root mean square residual, and common part accounted for) which are plotted against the degrees of freedom in a two-dimensional scree-like plot. The Hull method seeks an optimal balance between the fit and the degrees of freedom iteratively. The number of dimensions is determined, in which the elbow is located heuristically on a break or discontinuity in the convex hull. Lorenzo-Seva et al. (2011) conducted simulation studies on continuous variables to compare the Hull method and the PA method using the Bayesian information criterion, and observed that the Hull method outperforms the PA method in recovering the correct number of dimensions.

The DETECT method is a statistical method for dimensionality assessment. It aims to identify not only the number of dimensions, but also by which items a dimension is predominantly measured. The DETECT method is based on covariance theory that items measuring the same dimension would have positive covariances, whereas items measuring different dimensions would have negative covariances, given the latent traits or total scores have been taken into account (Stout et al. 1996; Zhang 2007; Zhang and Stout 1999). The DETECT index is expected to be zero if data are truly unidimensional. If data are multidimensional, the expected pairwise conditional covariance will be positive when both items measure the same dimension, and negative when they measure different dimensions. At the outset, a genetic algorithm, together with a hierarchical cluster analysis, can be used to search an optimal subspace. The optimal dimensionality partition is obtained by searching all over possible dimensions to maximize the DETECT index. The approximate simple structure index and the ratio index can help determine whether a dataset displays an approximately simple structure (in a simple structure, each item measures a single dimension, but different items may measure different dimensions).

The spectral clustering (SC) method aims to classify data structure into clusters on a manifold space (Luxburg 2007). It is widely used to cluster image data or pattern representation in computer science, statistics, etc. Moreover, it is easy to implement via standard linear algebra methods and usually outperforms traditional approaches such as the *k*-means clustering algorithm (Luxburg 2007). Once the number of clusters (dimensions) is determined, the SC method continues to classify

variables (items) to different clusters. Analogous to the DETECT method, the SC method identifies not only the number of clusters, but also the cluster to which an item belongs. Although the SC method is promising, to the authors' knowledge, it has never been applied to dimensionality assessment within the IRT framework. In this study, we thus evaluated the performance of an extended version of the SC method within the IRT framework. The details of the SC method and its extended version are given in the following section.

To sum up, this study aimed to compare the PA, DETECT, Hull, and SC methods in dimensionality assessment of IRT data through simulations. Most IRT models, including the one-, two-, and three-parameter logistic models for dichotomous items (Birnbaum 1968; Rasch 1960), and the partial credit model (Masters 1982), generalized partial credit model (Muraki 1992), and graded response model for polytomous items were examined. The rest of this article is organized as follows. First, we introduce the background of the SC method and its variant of spectral multi-manifold clustering (SMMC), which is rarely (or never) used in the IRT field. Second, we describe the IRT models that were used for data generation. Third, we summarize the results of the simulations that were conducted to evaluate the performance of the PA, DETECT, Hull, and SC methods in dimensionality assessment. Finally, we draw some conclusions and make suggestions for future studies.

Spectral Multi-manifold Clustering (SMMC)

In machine-learning terminology, a low-dimensional manifold is a topological space hidden in a high-dimensional data space. Conceptually, each item is located in an N-dimensional space, and items measuring the same construct should form a one-dimensional manifold. Trying to discover the dimensionality of item response data is intuitively analogous to discovering their latent manifolds. The characteristics of the *similar* items are learned by a machine-learning algorithm based on the responses. The properties of a high-dimensional data space are preserved in a low-dimensional embedding of the data, a process known as *dimensionality reduction*. After feature extraction, the next step is to classify items into appropriate clusters in the low-dimensional data space. The two-step processes are carried out consecutively in the SC method.

In essence, the SC method starts with generating an undirected affinity matrix $\mathbf{G}(V, E)$, which is derived from pairwise similarities (e.g., Euclidean distance). The term "undirected" means an equal weight between two data points, regardless of their direction. V and E stand for vertices (i.e., data points) and edge (i.e., the weight distance between vertices), respectively. The affinity matrix can be constructed by the ε-neighborhood graph, the k-nearest neighbor graphs, or the fully connected graph. The choice of construction depends on the problems to be solved. First, one can select the k-nearest neighbor graphs where a data point connects to only k-nearest data points near itself, finally resulting in a sparse, undirected, and

weighted adjacency matrix **W** (Luxburg 2007). Second, the normalized SC algorithm can be used (Shi and Malik 2000), which meets the objectives of clustering (i.e., minimize the between-cluster similarities and maximize the within-cluster similarities), consistence (i.e., the clustering results converge to a true partition when data increase), and simple computation (Luxburg 2007). Therefore, the data points can be mapped into a low-dimensional space, where the characteristics of sets of items become more evident than in the original space (i.e., similar items are close to one another than to other items). Finally, the number of dimensions is determined by counting the first k smallest eigenvalues equal to zero, and then the traditional k-means algorithm can be used to cluster an eigenvector matrix derived from a generalized eigenproblem. Items within the same cluster are deemed as measuring the same dimension.

The normalized SC algorithm is outlined as follows. First, a similarity graph is constructed using the k-nearest neighbor method (or the other methods outlined above). Second, an unnormalized graph Laplacian matrix is computed, which is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where **D** is a degree matrix, defined as the diagonal matrix with the degrees $d_1, \ldots, d_I$ on the diagonal. The d_i is defined as. Third, a generalized eigenproblem $\mathbf{L}u = \lambda \mathbf{D}u$ is solved to derive the first k generalized eigenvector $u_1, \ldots, u_k$ corresponding to the k smallest eigenvalues, then a matrix $\mathbf{U} \in \mathbb{R}^{I \times K}$ is obtained, where K is the number of intrinsic dimension. The element of **U** is $y_i \in \mathbb{R}^K$ for $i = 1, \ldots, I$. Fourth, the items are grouped by the k-means algorithm into clusters.

In a multidimensional space, we conjecture that items in the same group (i.e., measuring the same dimension) may intersect with, or be nearly close to, other groups of items, especially when these dimensions are highly correlated. However, the normalized SC algorithm merely includes a local similarity among points in the neighborhood. It may not be informative to accurately cluster items. Thus, the structural similarity (i.e., similar local tangent space) is another source of information derived from the data. Wang et al. (2011) proposed the SMMC algorithm to consider the information of local similarity and structural similarity to compositely create an affinity matrix **W**. The main idea is that if two items are close to each other and have similar tangent spaces, they may come from the same manifold (dimension). Then, local tangent space i at an item i can be approximated mixtures of probabilistic principal component analyzers (Tipping and Bishop 1999). The tangent space of pairwise points can be defined as, where h is the number of dimensions of the manifolds; ω is a weighting parameter; and form the principal angles between two tangent spaces and. The is defined as where, $l = 2, \ldots, h$.

The local similarity s_{ij} is defined as

$$s_{ij} = \begin{cases} 1 & \text{if } x_i \in Knn(x_j) \text{ or } x_j \in Knn(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $Knn(x)$ denotes k-nearest neighbors of x. Finally, the local similarity function and structural similarity function are integrated to generate the affinity value:

$$w_{ij} = s_{ij}t_{ij} = \begin{cases} \left[\sum_{l=1}^{h} \cos(\phi_l)\right]^{\omega} & \text{if } x_i \in Knn(x_j) \text{ or } x_j \in Knn(x_i) \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

Each item response vector is transformed into a standardized score to remove the impact of different variances of response vector in calculating the Euclidean distance between items.

Overall, the SMMC method uses only the distance information of local similarity function and structural similarity of the items, and does not assume the underlying item response function. More item categories could reduce the measurement error in calculating the Euclidean distances between items (i.e., the ordinal scale could be approximated as an interval scale as the number of categories increases), which could increase the accuracy. Sample size could also be an important factor and were investigated in the simulation studies.

Multidimensional IRT Models

The multidimensional generalized partial credit model (MGPCM) (Yao and Schwarz 2006) was used to generate item responses. It is commonly used for achievement tests in the psychometric literature. The item response function of the MGPCM is defined as

$$\Pr(Z_{ni} = z|\theta_n, \xi_i) = \frac{\exp[z(\alpha_i^T\theta_n - \delta_i) - \Sigma_{k=0}^{z}\tau_{ik}}{\Sigma_{m=0}^{C-1}\left\{\exp[m(\alpha_i^T\theta_n - \delta_i) - \Sigma_{k=0}^{m}\tau_{ik}]\right\}}, \quad (3)$$

where $Z_i \in [0, 1, \ldots, C-1]$ is the observable response variable; z is the observed response; C is the number of categories; $\boldsymbol{\theta}_n$ is the latent trait vector representing $D \times 1$ dimensions for person, n; ξ_i is item i's parameter vector containing a $D \times 1$ slope parameter vector of $\boldsymbol{\alpha}_i$, an intercept parameter of δ_i; and threshold parameters of $\tau_{i0}, \ldots, \tau_{iC-1}$. τ_{i0} is set at 0 by convention. The MGPCM is compensatory because the latent traits $\boldsymbol{\theta}$ are connected by a linear combination of $\boldsymbol{\alpha}_i$, when an item measures multiple latent traits. Dimensionality is the minimum number of latent traits that is adequate to explain the underlying examinees' performance (assuming local independence and monotonicity) and groups of items being sensitive to differences along the dimensions (Reckase 2009; Svetina and Levy 2014).

To be in line with most literature on dimensionality assessment, we focus on simple structures of latent traits in this study. That is, each item measures a single dimension and different items may measure different latent traits. Simple structures are also referred to as between-item multidimensionality, in contract to within-item multidimensionality where an item may measure multiple latent traits simultaneously (Adams et al. 1997).

Simulations

The SMMC, PA, and Hull methods have been implemented and are available at http://lamda.nju.edu.cn/code_SMMC.ashx (Wang et al. 2011) and http://psico.fcep.urv.es/utilitats/factor/, respectively, where the last two methods have been implemented in a stand-alone program called FACTOR 8.1 (Lorenzo-Seva and Ferrando 2006). The DOS version of the poly-DETECT program is generously provided by the author (Zhang 2007).

In this study, the independent variables included: (a) number of respondents (N = 250; 1000; 2000), (b) number of response categories (C = 2, 4, 6), (c) number of dimensions (D = 1, 2, 3), (d) correlation among dimensions (r = 0, 0.4, 0.8), and (e) number of items per dimension (L = 10, 20). Each dimension had the same number of items in order to obtain a constant amount of information across dimensions, so that the effect of the number of items would not be confounded with that of the number of dimensions. This setting was also adopted in the literature (Garrido et al. 2011; Svetina 2012). A total of 100 replications in each condition were implemented. For each simulated dataset, the four methods (PA, Hull, DETECT, and SMMC) were adopted, and their performance in detecting the correct dimensionality was compared.

The dependent variable was the accuracy rate, defined as the proportion of times in the 100 replications that the number of dimensions was identified correctly. Additionally, for the SMMC and DETECT methods, it was interesting to know how accurate an item was assigned to its corresponding dimension, given that the number of dimensions had already been identified correctly. The following hit rate was proposed to reflect the accuracy:

$$\frac{\sum_{r=1}^{\mathrm{AR}} \sum_{i=1}^{I} h_{ri}}{\mathrm{AR} \times I}, \tag{4}$$

where AR was the number of times that the number of dimensions was correctly identified in 100 replications; h_{ri} was a dummy variable and was equal to 1 if an item was correctly assigned to its dimension, and 0 otherwise; and I was the test length. When the hit rate was 1, all items were correctly assigned to their dimensions; when the hit rate was 0, none of the items was correctly assigned.

Parameter configurations were as follows. In the SMMC method, the dimension of the manifolds was set at 1, meaning that each manifold was assumed to be unidimensional; the number of mixtures was set at 2 due to a relatively small number of items; the number of neighbors was set at 2; and the weighting parameter was arbitrarily set at 10. The number of clusters was determined by the eigengap heuristic (Luxburg 2007), which counts the number of times from the first eigenvalue to the last one that the difference in two successive eigenvalues was relatively

smaller than 10^{-6}. As for the PA method, a total of 100 random datasets were generated, and their Pearson correlation matrices were calculated. Factors were extracted using the unweighted least squares with the Promin oblique rotation. Finally, the 95th percentile criterion was used to determine the number of dimensions. The same factor extraction method as in the PA method was adopted for the Hull method, and the comparative fit index was used to describe model-data fit in the Hull method. Regarding the DETECT method, an exploratory approach was adopted, because no prior information about the dimensionality was utilized in the simulations. The minimum number of respondents per score strata was set at 5, meaning that the strata with less than 5 respondents would be removed from the calculation of conditional covariances and then collapsed into adjacent strata. The number of mutations in the genetic algorithm was set at 2 for 10-item tests and 4 for 20-item tests. The maximum number of dimensions for the exploratory search was set at 6. Cross-validation, strongly recommended in small sample sizes or short tests (Monahan et al. 2007), was utilized, in which the respondents were randomly split into two halves to serve as the training and validation subsets. To examine the dimensional structure, the critical values for approximate simple structure index (ASSI) and ratio index (R) were set at 0.25 and 0.36, respectively, which were the default values in poly-DETECT program. When they both were smaller than the critical values, the dataset would be declared as approximate simple structure. The expected conditional covariance was calculated for the affinity matrix.

Only simple structures were considered in this study. In simulating item responses, the discriminate parameter vector $\boldsymbol{\alpha}_i$ was set at $\mathbf{1}$ for the MGPCM; the person parameter vector $\boldsymbol{\theta}_n$ was generated from a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance-variance matrix $\boldsymbol{\Sigma}$ with diagonal elements equal to 1. The correlations among dimensions were all set at a constant. To be consistent with most IRT literature, the person parameters were treated as random effects, but the item parameters as fixed effects, across replications. The intercept parameter δ was set from -2 to 2 with an equal interval between two adjacent items for each dimension. The settings of item and person parameters were generally consistent with simulation and empirical studies in the IRT literature (Muraki 1992). For example, Muraki (1992) uses the range from -1.68 to 1.68 in simulations and obtains a range from -4.47 to 3.37 in an empirical example.

The $\tau_{i1}, \ldots, \tau_{iC}$ were drawn from a uniform interval from -2 to 2 with the order of $\tau_{i1} < \tau_{i2} < \ldots < \tau_{i(C-1)} < \tau_{iC}$.

We had the following expectations on the simulation results. The PA, Hull, and DETECT methods would be superior to the SMMC method, because the former three methods considered the ordinal nature of categorical data, whereas the latter was a nonparametric method and only consider the continuous nature. The SMMC method might outperform the others as the number of categories is large.

Results

Dichotomous Items

Table 1 shows the accuracy rates and the mean numbers of misspecified dimensions (in parentheses) for two-category items following the one-, two-, and three-dimensional generalized partial credit models. First, consider one-dimensional data. The PA and Hull methods performed almost perfectly across all conditions. The DETECT method did not yield results when the sample size $N = 250$ (DETECT program warned the calculated conditional covariances was inaccurate and did not give results), but yielded good accuracy rates when $N = 1000$ or 2000. The SMMC method performed satisfactorily only when $N = 1000$ or 2000, and test length $L = 20$ items (but sometimes, it failed to converge).

Next, consider two- and three-dimensional data. When $N = 250$ and $r = 0$ or 0.4, the PA and Hull methods outperformed the other two methods in the accuracy rates; when $N = 250$ and $r = 0.8$, the SMMC method performed relatively better. The DETECT method was the best when $N = 1000$ or 2000. The hit rates were perfect or nearly perfect for the DETECT method and very high for the SMMC method. Generally, it was much more difficult to identify dimensionality when $r = 0.8$ than when $r = 0.4$ or 0. Fortunately, the DETECT method still yielded a very high accuracy rate when $r = 0.8$, given that $N = 2000$ and $L = 20$.

Polytomous Items

Tables 2 and 3 show the accuracy rates and the mean numbers of misspecified dimensions (in parentheses) for four-category items and six-category items following the one-, two-, and three-dimensional generalized partial credit models, respectively. A comparison of Tables 1 (two-category items), 2 (four-category items), and 3 (six-category items) revealed that the more response categories, the easier the identification of dimensionality. Given the great similarity between Tables 2 and 3, the following discussion focuses on Table 3. First, consider one-dimensional data. The PA method performed perfectly in the identification of the single dimension and outperformed the other three methods. The DETECT method did not yield results when $N = 250$ but performed almost perfectly when $N = 1000$ or 2000. The Hull method performed perfectly when $L = 10$ but very poorly when $L = 20$ and $N = 1000$ or 2000. The SMMC method performed satisfactorily when $N = 1000$ or 2000 (but sometimes, it failed to converge).

Next, consider two- and three-dimensional data. When $N = 250$ and $r = 0$ or 0.4, the PA, Hull, and SMMC methods performed almost perfectly in the identification of the two or three dimensions, whereas the DETECT method did not yield results. When $N = 250$ and $r = 0.8$, the SMMC method was the best. When $N = 1000$ or 2000, the DETECT method always yielded a perfect accuracy rate. The PA method

Table 1 Accuracy rates, mean numbers of misspecified dimensions (in parentheses), and hit rates (DETECT/SMMC) for two-category items following the one-, two-, and three-dimensional generalized partial credit models

1D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
PA	1			1			0.99 (2)		
Hull	0.99 (2)			1			1		
DETECT	NA			0.86 (NA)			0.59 (NA)		
SMMC	0.05 (2.2)			0.79 (2)			0.80 (NA)		
1D/20 items	250			1000			2000		
PA	1			0.98 (2)			0.99 (2)		
Hull	1			1			1		
DETECT	NA			1			1		
SMMC	0.06 (2.6)			0.96 (NA)			0.88 (NA)		
2D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	0.78 (1)	0.01 (1)	1	0.72 (1)	0.01 (1)	1	0.59 (1)	0.01 (1)
Hull	0.98 (3)	0.74 (1.2)	0 (1)	1	0.99 (1)	0 (1)	1	1	0.01 (1)
DETECT	NA	NA	NA	1	1	0.03 (1.1)	1	1	0.36 (1.9)
SMMC	0.75 (1.5)	0.45 (2.3)	0.43 (3.3)	0.99 (3)	0.90 (1.2)	0.01 (1)	1	0.98 (NA)	0.02 (1)
Hit rate	NA/0.99	NA/0.87	NA/0.57	1/1	1/1	0.96/0.65	1/1	1/1	1/1
2D/20 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	1	0.05 (1)	1	1	0.03 (1)	1	1	0.01 (1)
Hull	1	0.98 (1)	0 (1)	1	1	0 (1)	1	1	0 (1)

(continued)

Table 1 (continued)

1D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
DETECT	NA	NA	NA	1	1	0.13 (1.3)	1	1	0.98 (3)
SMMC	0.90 (1.4)	0.54 (1.7)	0.15 (3.9)	1	0.91 (1)	0.02 (1)	1	1	0.02 (1)
Hit rate	NA/1.00	NA/0.95	NA/0.55	1/1	1/1	0.97/0.60	1/1	1/1	1.00/1
3D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	0.99 (4)	0.63 (1.7)	0 (1)	0.99 (4)	0.81 (1.8)	0 (1)	0.99 (4)	0.75(1.7)	0 (1.0)
Hull	1	0.38 (1.1)	0 (1)	1	0.99 (1)	0 (1)	1	1	0 (1)
DETECT	NA	NA	NA	1	1	0.01 (1.1)	1	1	0.53 (1.6)
SMMC	0.34 (3.5)	0.30 (3.6)	0.32 (3.8)	0.99 (4)	0.67 (1.9)	0 (1.0)	1	1	0 (1.0)
Hit rate	NA/0.87	NA/0.61	NA/0.38	1/1	1.00/1.00	0.78/NA	1/1	1/1	0.95/NA
3D/20 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	1	0 (1)	1	1	0 (1.1)	1	0.99 (4)	0 (1.1)
Hull	1	0.74 (1)	0 (1)	1	1	0 (1)	1	1	0 (1)
DETECT	NA	NA	NA	1	1	0.04 (1)	1	1	1
SMMC	0.45 (2.2)	0.19 (2.8)	0.19 (4.8)	1	0.75 (2)	0 (1)	1	0.99 (4)	0 (1)
Hit rate	NA/0.99	NA/0.70	NA/0.42	1/1	1/1	0.90/NA	1/1	1/1	0.99/NA

Note NA not available; hit rates are for the DETECT and SMMC methods

Table 2 Accuracy rates, mean numbers of misspecified dimensions (in parentheses), and hit rates (DETECT/SMMC) for four-category items following the one-, two-, and three-dimensional generalized partial credit models

1D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
PA	1			1			1		
Hull	0.92 (2)			1			1		
DETECT	NA			0.96 (2)			0.97 (2)		
SMMC	0.17 (2.1)			0.74 (2)			0.74 (NA)		
1D/20 items	250			1000			2000		
PA	1			1			1		
Hull	0.92 (2)			0.39 (2)			0.05 (2)		
DETECT	NA			0.99 (NA)			1		
SMMC	0.25 (2.3)			0.90 (2)			0.86 (2)		
2D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	1	0 (1)	1	1	0 (1)	1	1	0 (1)
Hull	1	0.99 (3)	0 (1)	1	1	0 (1.3)	1	1	0 (1.5)
DETECT	NA	NA	NA	1	1	1	1	1	1
SMMC	0.97 (3)	0.93 (2.6)	0.55 (2.3)	0.96 (3)	0.94 (3)	0.78 (1)	0.98 (NA)	0.92 (NA)	0.92 (1)
Hit rate	NA/1	NA/1	NA/0.72	1/1	1/1	1.00/1	1/1	1/1	1/1
2D/20 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	1	0.72 (1.7)	1	1	0.82 (3)	1	1	0.91 (1.7)
Hull	1	0.99 (3)	0 (1.1)	0.96 (4)	0.99 (4)	0 (1.5)	0.75 (4)	0.61 (4)	0 (1.6)

(continued)

Table 2 (continued)

1D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
DETECT	NA	NA	NA	1	1	1	1	1	1
SMMC	0.99 (3)	0.99 (4)	0.37 (1.5)	0.99 (NA)	0.98 (NA)	0.84 (1)	0.99 (3)	0.98 (NA)	0.94 (3)
Hit rate	NA/1	NA/1.00	NA/0.77	1/1	1/1	1/1	1/1	1/1	1/1
3D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	1	0 (1)	1	1	0 (1.0)	1	1	0 (1)
Hull	1	0.99 (1)	0 (1)	1	1	0 (1.1)	1	1	0 (1.1)
DETECT	NA	NA	NA	1	1	1	1	1	1
SMMC	0.96 (4)	0.84 (4.3)	0.27 (2.7)	0.98 (4)	0.99 (NA)	0.41 (1.7)	1	0.98 (NA)	0.91 (NA)
Hit rate	NA/1	NA/1.00	NA/0.57	1/1	1/1	1.00/1	1/1	1/1	1/1
3D/20 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	1	0.48(3.1)	1	1	0.05(4.0)	1	0.99(4)	0.06(4.0)
Hull	0.99 (4)	1	0 (1)	1	1	0 (1.1)	1	0.96 (6)	0 (1.2)
DETECT	NA	NA	NA	1	1	1	1	1	1
SMMC	0.98 (4)	0.98 (3)	0.29 (2.6)	1	0.98 (NA)	0.61 (1.9)	0.97 (4)	0.98 (4)	0.93 (3.3)
Hit rate	NA/1	NA/1	NA/0.51	1/1	1/1	1/1	1.00/1	1/1	1/1

Note NA not available; hit rates are for the DETECT and SMMC methods

Table 3 Accuracy rates, mean numbers of misspecified dimensions (in parentheses), and hit rates (DETECT/SMMC) for six-category items following the one-, two-, and three-dimensional generalized partial credit models

1D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
PA	1			1			1		
Hull	1			1			1		
DETECT	NA			0.94 (2)			0.95 (2)		
SMMC	0.18 (2.0)			0.78 (NA)			0.77 (NA)		
1D/20 items	250			1000			2000		
PA	1			1			1		
Hull	0.76 (2)			0.04 (2)			0 (2)		
DETECT	NA			1			1		
SMMC	0.22 (2.1)			0.88 (2)			0.91 (NA)		
2D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	1	0 (1)	1	1	0 (1)	1	1	0.01 (1)
Hull	1	1	0 (1)	1	1	0 (1.6)	1	1	0 (1.8)
DETECT	NA	NA	NA	1	1	1	1	1	1
SMMC	0.98 (3)	0.96 (3)	0.55 (1.8)	0.97 (3)	0.97 (NA)	0.95 (3)	0.98 (NA)	0.91 (3)	0.92 (3)
Hit rate	NA/1	NA/1	NA/0.87	1/1	1/1	1/1	1/1	1/1	1/1
2D/20 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	0.99 (3)	0.57 (2.9)	1	1	0.40 (3)	1	1	0.69 (3)
Hull	1	0.99 (3)	0 (1.3)	0.53 (4)	0.89 (4)	0 (1.9)	0.08 (4)	0.29 (4)	0 (2)

(continued)

Table 3 (continued)

1D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
DETECT	NA	NA	NA	1	1	1	1	1	1
SMMC	0.97 (3)	0.97 (3.3)	0.40 (1.3)	0.96 (3)	0.94 (3)	0.92 (2.6)	0.98 (3)	0.95 (3)	0.95 (NA)
Hit Rate	NA/1	NA/1	NA/0.96	1/1	1/1	1/1	1/1	1/1	1/1
3D/10 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	1	0 (1.0)	1	1	0 (1.0)	1	1	0 (1)
Hull	1	1	0 (1)	1	1	0 (1.3)	1	1	0 (1.6)
DETECT	NA	NA	NA	1	1	1	1	1	1
SMMC	0.98 (4)	0.97 (4)	0.31 (3.1)	0.95 (4)	0.96 (4)	0.92 (2.8)	1	0.93 (4)	0.95 (NA)
Hit rate	NA/1	NA/1	NA/0.72	1/1	1/1	1/1	1/1	1/1	1/1
3D/20 items	$N = 250$			$N = 1000$			$N = 2000$		
	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$	$r = 0$	$r = 0.4$	$r = 0.8$
PA	1	0.99 (4)	0.23 (4.0)	1	1	0 (4.0)	1	1	0 (4.0)
Hull	1	1	0 (1.1)	0.97 (6)	1	0 (1.7)	0.26 (6)	0.94 (6)	0 (1.8)
DETECT	NA	NA	NA	1	1	1	1	1	1
SMMC	0.97 (4)	0.92 (4)	0.32 (3.2)	0.92 (4)	0.91 (4)	0.92 (3.6)	0.92 (4)	0.93 (4)	0.91 (4)
Hit rate	NA/1	NA/1	NA/0.81	1/1	1/1	1/1	1/1	1/1	1/1

Note NA not available; hit rates are for the DETECT and SMMC methods

always yielded a perfect accuracy rate when $r = 0$ or 0.4, but it performed very poorly when $r = 0.8$. The Hull method performed poorly when $L = 20$ or $r = 0.8$. The SMMC performed almost perfectly across all conditions. The hit rates were always perfect for the DETECT method, except when the sample size was 250, and often perfect for the SMMC method. In summary, the following recommendations appear applicable:

1. The PA method is recommended when sample sizes are small, and the number of dimension is one or the correlations among dimensions are low.
2. The DETECT method is recommended when sample sizes are large.
3. The SMMC method can be a supplement to the PA and DETECT methods when the number of categories is large.

Conclusion and Discussion

The SMMC method makes no assumption on the responding process, seeming to be a promising alternative for dimensionality assessment. Although the SMMC performs worse than other methods in general, it can serve as a supplement to the PA and DETECT methods.

Future studies can aim at evaluating these four methods and other methods under a more comprehensive design. For example, multidimensional scaling and bootstrap generalization are promising methods of dimensionality assessment (Finch and Monahan 2008; Meara et al. 2000). In this study, only simple structures were investigated. The dimensionality assessment of complex structures, in which an item may measure more than one dimension and the dimensions can be compensatory or noncompensatory (Embretson 1997; Sympso 1978; Whitely 1980), is of great importance and left for future studies. The PA method used in this study adopted the Pearson correlation in order to be consistent with the literature (Timmerman and Lorenzo-Seva 2011; Weng and Cheng 2005). Actually, the polychoric correlation can be used for ordinal data (Cho et al. 2009; Timmerman and Lorenzo-Seva 2011).

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

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 (Chaps. 17–20)*. Reading, MA: Addison-Wesley.

Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement, 69*(5), 748–759.

Embretson, S. E. (1997). Multicomponent response models. *Handbook of modern item response theory* (pp. 305–321), Springer.

Finch, H., & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. *Applied Measurement in Education, 21*(2), 119–140.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement, 71*(3), 551–570.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). Factor: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, Instruments, & Computers, 38*, 88–91.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46*(2), 340–364.

Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395–416.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Meara, K., Robin, F., & Sireci, S. G. (2000). Using multidimensional scaling to assess the dimensionality of dichotomous item data. *Multivariate Behavioral Research, 35*(2), 229–259.

Monahan, P. O., Stump, T. E., Finch, H., & Hambleton, R. K. (2007). Bias of exploratory and cross-validated DETECT index under unidimensionality. *Applied Psychological Measurement, 31*(6), 483–503.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedogogiske Institut, 1960. Chicago: University of Chicago Press, 1980.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer-Verlag.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888–905.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, Jinming. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331–354.

Svetina, D. (2012). Assessing dimensionality of noncompensatory multidimensional item response theory with complex structures. *Educational and Psychological Measurement*.

Svetina, D., & Levy, R. (2014). A framework for dimensionality assessment for multidimensional item response models. *Educational Assessment, 19*(1), 35–57.

Sympson, J. B. (1978). *A model for testing with multidimensional items.* Paper presented at the Proceedings of the 1977 computerized adaptive testing conference.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209–220.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61*(3), 611–622.

Tran, U. S., & Formann, A. K. (2009). Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement, 69*(1), 50–61.

Wang, Y., Jiang, Y., Wu, Y., & Zhou, Z.-H. (2011). Spectral clustering on multiple manifolds. *IEEE Transactions on Neural Networks, 22*(7), 1149–1161.

Weng, L.-J., & Cheng, C.-P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65*(5), 697–716.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*(4), 479–494.

Wilson, D. T., Wood, R., & Gibbons, R. D. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*. SSI, Scientific Software International.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests. *Applied Psychological Measurement, 30*(6), 469–492.

Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika, 72*(1), 69–91.

Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213–249.

The Lexile Framework for Reading: An Introduction to What It Is and How to Use It

Malbert Smith, Jason Turner, Eleanor Sanford-Moore and Heather H. Koons

Introduction

The Lexile Framework for Reading was developed over a ten-year period beginning in 1984. The research that provided the basis for the framework was motivated by the belief that reading scores from an absolute scale would be a meaningful and useful addition to the kinds of norm-referenced data typically provided by traditional reading comprehension tests, which compare students in relative terms. Too often, students received a score on a reading comprehension test that could not be compared to scores from other reading comprehension tests the student had taken. Additionally, reading scores frequently had little utility. Because both reader and text measures are on the same scale, Lexile measures provide information that can be used to guide readers in their reading selections, offering guidance to readers who want to increase the challenge of the materials they can read and comprehend.

Background

A reader's comprehension of text is dependent on many factors—the purpose for reading, the ability of the reader, and the text that is being read. The reader can be asked to read a text for many purposes including entertainment (literary experience), to gain information, or to perform a task. Each reader brings to the reading experience a variety of important factors: reading ability, prior knowledge, interest level, and developmental readiness. For any text, there are three factors associated with the readability of the text: complexity, support, and quality. All of these reader and text factors are important considerations when evaluating the appropriateness of

M. Smith · J. Turner · E. Sanford-Moore (✉) · H.H. Koons
MetaMetrics, Inc., Durham, NC, USA
e-mail: esanford@lexile.com

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_27

a text for a reader. The Lexile Framework focuses primarily on two features: reader ability and text complexity.

All symbol systems share two features: a semantic component and a syntactic component. In language, the semantic units are words. Words are organized according to rules of syntax into thought units and sentences (Carver 1974). In all cases, the semantic units vary in familiarity and the syntactic structures vary in complexity. The comprehensibility (or difficulty) of a message is dominated by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message. The Lexile Framework utilizes these two dominant features of language in measuring text complexity by examining the characteristics of word frequency and sentence length.

The Semantic Component

Most operationalizations of semantic complexity are proxies for the probability that an individual will encounter a word in a familiar context and thus be able to infer its meaning (Bormuth 1966). This is the basis of exposure theory, which explains the way receptive or hearing vocabulary develops (Miller and Gildea 1987; Stenner et al. 1983). Klare (1963) hypothesized that the semantic component varied along a familiarity-to-rarity continuum. This concept was further developed by Carroll et al. (1971), whose word frequency study examined the reoccurrence of words in a five-million-word corpus of running text. Knowing the frequency of words as they are used in written and oral communication provided the best means of inferring the likelihood that a word would be encountered by a reader and thus become a part of that individual's receptive vocabulary.

In a study examining receptive vocabulary, Stenner et al. (1983) analyzed more than 50 semantic variables in order to identify those elements that contributed to the difficulty of the 350 vocabulary items on Forms L and M of the *Peabody Picture Vocabulary Test*—Revised (Dunn and Dunn 1981). Variables included part of speech, number of letters, number of syllables, the modal grade at which the word appeared in school materials, content classification of the word, the frequency of the word from two different word counts, and various algebraic transformations of these measures.

The word frequency measure used was the raw count of how often a given word appeared in a corpus of 5,088,721 words sampled from a broad range of school materials (Carroll et al. 1971). A "word family" included. (1) the stimulus word; (2) all plurals (adding "-s" or changing "-y" to "-ies"); (3) adverbial forms; (4) comparatives and superlatives; (5) verb forms ("-s," "-d," "-ed," and "-ing"); (6) past participles; and (7) adjective forms. Correlations were computed between algebraic transformations of these means and the rank order of the test items. Since the items were ordered according to increasing difficulty, the rank order was used as the observed item difficulty. The mean log word frequency provided the highest correlation with item rank order ($r = -0.779$) for the items on the combined form.

The Lexile Framework currently employs a 600-million-word corpus when examining the semantic component of text. This corpus was assembled from more than 15,000 texts that were measured by MetaMetrics for publishers from 1998 to 2002. When text is analyzed by MetaMetrics, all electronic files are initially edited according to established guidelines used with the Lexile Analyzer software. These guidelines include the removal of all incomplete sentences, chapter titles, and paragraph headings; running of a spell check; and re-punctuating where necessary to correspond to how the book would be read by a child (for example, at the end of a page). The text is then submitted to the Lexile Analyzer that examines the lengths of the sentences and the frequencies of the words and reports a Lexile measure for the book. When enough additional texts have been analyzed to make an adjustment to the corpus necessary and desirable, a linking study will be conducted to adjust the calibration equation such that the Lexile measure of a text based on the current corpus will be equivalent to the Lexile measure based on the new corpus.

The Syntactic Component

Klare (1963) provided a possible interpretation for how sentence length works in predicting passage difficulty. He speculated that the syntactic component varied with the load placed on short-term memory. Crain and Shankweiler (1988), Shankweiler and Crain (1986), and Liberman et al. (1982) have also supported this explanation. The work of these individuals has provided evidence that sentence length is a good proxy for the demand that structural complexity places upon verbal short-term memory.

While sentence length has been shown to be a powerful proxy for the syntactic complexity of a passage, an important caveat is that sentence length is not the underlying causal influence (Chall 1988). Researchers sometimes incorrectly assume that manipulation of sentence length will have a predictable effect on passage difficulty. Davidson and Kantor (1982), for example, illustrated rather clearly that sentence length can be reduced and difficulty increased and vice versa.

Based on previous research, sentence length was selected as a proxy for the syntactic component of reading complexity in the Lexile Framework.

Calibration of Text Complexity

A research study on semantic units conducted by Stenner et al. (1983) was extended to examine the relationship of word frequency and sentence length to reading comprehension. In Stenner et al. (1987) performed exploratory regression analyses to test the explanatory power of these variables. This analysis involved calculating the mean word frequency and the log of the mean sentence length for each of the 66 reading comprehension passages on the *Peabody Individual Achievement Test*.

The observed difficulty of each passage was the mean difficulty of the items associated with the passage (provided by the publisher) converted to the logit scale. A regression analysis based on the word frequency and sentence length measures produced a regression equation that explained most of the variance found in the set of reading comprehension tasks. The resulting correlation between the observed logit difficulties and the theoretical calibrations was 0.97 after correction for range restriction and measurement error. The regression equation was further refined based on its use in predicting the observed difficulty of the reading comprehension passages on eight other standardized tests. The resulting correlation between the observed logit difficulties and the theoretical calibrations when the nine tests were combined into one was 0.93 after correction for range restriction and measurement error.

Once a regression equation was established linking the syntactic and semantic features of text to the complexity of text, the equation were used to calibrate test items and text.

The Lexile Scale

In developing the Lexile scale, the Rasch item response theory model (Wright and Stone 1979) was used to estimate the difficulties of items and the abilities of persons on the logit scale. The calibrations of the items from the Rasch model are objective in the sense that the relative difficulties of the items will remain the same across different samples of persons (specific objectivity). When two items are administered to the same person, it can be determined which item is harder and which one is easier. This ordering is likely to hold when the same two items are administered to a second person. If two different items are administered to the second person, there is no way to know which set of items is harder and which set is easier. The problem is that the location of the scale is not known. General objectivity requires that scores obtained from different test administrations be tied to a common zero—absolute location must be sample independent (Stenner 1990). To achieve general objectivity, the theoretical logit difficulties must be transformed to a scale where the ambiguity regarding the location of zero is resolved.

The first step in developing a scale with a fixed zero was to identify two anchor points for the scale. The following criteria were used to select the two anchor points: they should be intuitive, easily reproduced, and widely recognized. For example, most thermometers have anchor points at the freezing and boiling points of water. For the Lexile scale, the anchor points are text from seven basal primers for the low end and text from *The Electronic Encyclopedia* (Grolier Inc. 1986) for the high end. These points correspond to the middle of first-grade text and the midpoint of workplace text.

The next step was to determine the unit size for the scale. For the Celsius thermometer, the unit size (a degree) is 1/100th of the difference between freezing (0°) and boiling (100°) water. For the Lexile scale, the unit size was defined as

1/1000th of the difference between the mean difficulty of the primer material and the mean difficulty of the encyclopedia samples. Therefore, a Lexile unit by definition equals 1/1000th of the difference between the comprehensibility of the primers and the comprehensibility of the encyclopedia.

The third step was to assign a value to the lower anchor point. The low-end anchor on the Lexile scale was assigned a value of 200.

Finally, a linear equation of the form

$$[(\text{Logit} + \text{constant}) \times \text{CF}] + 200 = \text{Lexile text measure} \quad (1)$$

was developed to convert logit difficulties to Lexile calibrations. The values of the conversion factor (CF) and the constant were determined by substituting in the anchor points and then solving the system of equations.

The Lexile Scale ranges from below 200L to above 1600L. There is a not an explicit bottom or top to the scale, but rather two anchor points on the scale (described above) that describe different levels of reading comprehension. The Lexile Map, a graphic representation of the Lexile Scale from 200L to 1600L, provides a context for understanding reading comprehension.

Forecasting Comprehension with the Lexile Framework

A reader with a measure of 600L who is given a text measured at 600L is expected to have a 75 % comprehension rate. This 75 % comprehension rate is the basis for selecting text that is targeted to a reader's reading ability, but what exactly does it mean? And what would the comprehension rate be if this same reader were given a text measured at 350L or one at 850L?

The 75 % comprehension rate for a reader-text pairing can be given an operational meaning by imagining the text to be carved into item-sized slices of approximately 125–140 words with a question embedded in each slice. A reader who answers three-fourths of the questions correctly has a 75 % comprehension rate.

Suppose, instead that the text and reader measures are not the same. It is the difference in Lexile measures between reader and text that governs comprehension. If the text measure is less than the reader measure, the comprehension rate will exceed 75 %. If not, it will be less. The question is "By how much?" What is the expected comprehension rate when a 600L reader reads a 350L text?

If all the item-sized slices in the 350L text had the same calibration, the 250L difference between the 600L reader and the 350L text could be determined using the Rasch item response theory (IRT) model equation. This equation describes the relationship between the measure of a student's level of reading comprehension and the calibration of the items. Unfortunately, comprehension rates calculated by this procedure would be biased because the calibrations of the slices in ordinary prose are not all the same. The average difficulty level of the slices and their variability both affect the comprehension rate.

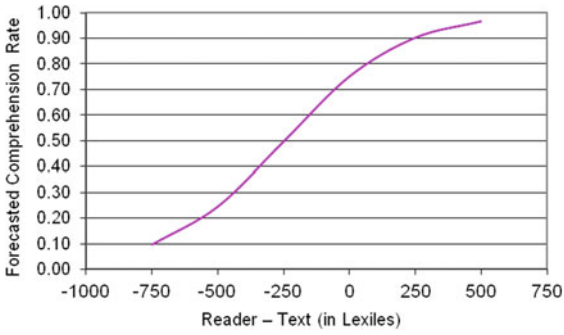


Fig. 1 Relationship between reader-text discrepancy and forecasted reading comprehension rate

Although the exact relationship between comprehension rate and the pattern of slice calibrations is complicated, Eq. 2 is an unbiased approximation.

$$Rate = \frac{e^{ELD + 1.1}}{1 + e^{ELD + 1.1}} \tag{2}$$

where ELD is the "effective logit difference" given by

$$\text{ELD} = (\text{Reader Lexile measure} - \text{Text Lexile measure}) \div 225 \tag{3}$$

Figure 1 shows the general relationship between reader-text discrepancy and forecasted comprehension rate. When the reader measure and the text measure are the same (difference of 0L on the *x*-axis), then the forecasted comprehension rate is 75 %. In the example in the preceding paragraph, the difference between the reader measure of 600L and the text measure of 350L is 250L. Referring to Fig. 1 and using +250L (reader minus text), the forecasted comprehension rate for this reader-text combination would be 90 %.

Tables 1 and 2 show comprehension rates calculated for various combinations of reader measures and text measures. As discussed later in this paper, the match of text with a known Lexile text measure to a reader with a known Lexile reader measure can be adjusted to meet the desired goals of the reader or instructional purposes.

Table 1 Comprehension rates for the same individual with materials of varying comprehension difficulty

Person measure	Text calibration	Sample titles	Forecast comprehension (%)
1000L	500L	Tornado (Byars)	96
1000L	750L	The martian chronicles (Bradbury)	90
1000L	1000L	Reader's digest	75
1000L	1250L	The call of the wild (London)	50
1000L	1500L	On the equality among mankind (Rousseau)	25

Table 2 Comprehension rates of differently abled persons with the same material

Person measure	Calibration for Sports Illustrated	Forecast comprehension (%)
500L	1000L	25
750L	1000L	50
1000L	1000L	75
1250L	1000L	90
1500L	1000L	96

The subjective experience of 50, 75, and 90 % comprehension as reported by readers varies greatly. A 1000L reader reading 1000L text (75 % comprehension) reports confidence and competence. Teachers listening to such a reader report that the reader can sustain the meaning thread of the text and can read with motivation and appropriate emotion and emphasis. In short, such readers sound like they comprehend what they are reading. A 1000L reader reading 1250L text (50 % comprehension) encounters so much unfamiliar vocabulary and difficult syntactic structures that the meaning thread is frequently lost. Such readers report frustration and seldom choose to read independently at this level of comprehension difficulty. Finally, a 1000L reader reading 750L text (90 % comprehension) reports control of the text, reads with speed, and experiences automaticity during the reading process.

The primary utility of the Lexile Framework is its ability to forecast what happens when readers confront text. With every application by teacher, student, librarian, or parent there is a test of the framework's accuracy. The Lexile Framework makes a point prediction every time a text is chosen for a reader. Anecdotal evidence suggests that the Lexile Framework predicts as intended. That is not to say that there is an absence of error in forecasted comprehension. There is error in text measures, reader measures, and their difference modeled as forecasted comprehension. However, the error is sufficiently small that the judgments about readers, texts, and comprehension rates are useful.

Using Lexile Measures

The Lexile Framework for Reading provides teachers and educators with tools to help them link assessment results with subsequent instruction. Assessments that are linked with the Lexile scale provide tools for monitoring the progress of students at any time during the course of instruction as well as important information for selecting appropriate reading material for students.

The Lexile Framework reporting scale is not bounded by grade level, although typical Lexile measure ranges have been identified for students in specific grades. Lexile measures do not translate specifically to grade levels. Within any grade, there will be a range of readers and a range of materials to be read. For example, in a fifth-grade classroom there will be some readers who are far ahead of the others and

there will be some readers who are behind the others in terms of reading ability. To say that some books are "just right" for fifth graders assumes that all fifth graders are reading at the same level. Because the Lexile Framework reporting scale is not bounded by grade level, it makes provisions for students who read below or beyond their grade level.

Simply because a student is an excellent reader, it should not be assumed that the student would necessarily comprehend a text typically found at a higher grade level. Without adequate background knowledge, the words may not have sufficient meaning to the student. A high Lexile measure for a grade indicates that the student can read grade- or age-appropriate materials at a high comprehension level.

Use the Lexile Framework to Select Books

Teachers, parents, and students can use the tools provided by the Lexile Framework to select materials to plan instruction. When teachers provide parents and students with lists of titles that match the students' Lexile measures, they can then work together to choose appropriate titles that also match the students' interests and background knowledge. *The Lexile Framework does not prescribe a reading program, but it gives educators more knowledge of the variables involved when they design reading instruction.* The Lexile Framework facilitates multiple opportunities for use in a variety of instructional activities. After becoming familiar with the Lexile Framework, teachers are likely to think of a variety of additional creative ways to use this tool to match students with books that students find challenging, but not frustrating.

Many factors affect the relationship between a reader and a book. These factors include text content, age of the reader, interests of the reader, suitability of the text, and text difficulty. The Lexile measure of a text, a measure of text complexity, is a good starting point in the selection process, but other factors also must be considered. The Lexile measure should never be the only piece of information used when selecting a text for a reader.

Teach Learning Strategies by Controlling Comprehension Match

The Lexile Framework permits the teacher to target readers with challenging text and to systematically adjust text targeting when the teacher wants fluency and automaticity (i.e., reader measure is well above text measure) or wants to teach strategies for attacking "hard" text (i.e., reader measure is below text measure). For example, metacognitive ability has been well documented to play an important role

in reading comprehension performance. Once teachers know the kinds of texts that would likely be challenging for a group of readers, they can systematically plan instruction that will allow students to encounter difficult text in a controlled fashion and make use of instructional scaffolding to build student success and confidence with more challenging text. The teacher can model appropriate learning strategies for students, such as rereading or rephrasing text in one's own words, so that students can then learn what to do when comprehension breaks down. Students can then practice these metacognitive strategies on selected text while the teacher monitors their progress.

Teachers can use Lexile measures to guide a struggling student toward texts at the lower end of the student's Lexile range (100L or more below to 50L above his or her Lexile measure). Similarly, advanced students can be adequately challenged by reading texts at the midpoint of their Lexile range, or slightly above. Challenging new topics or genres may be approached in the same way.

Differentiating instruction for the reading experience also involves the student's motivation and purpose. If a student is highly motivated for a particular reading task (e.g., self-selected free reading), the teacher may suggest books higher in the student's Lexile range. If the student is less motivated or intimidated by a reading task, material at the lower end of his or her Lexile range can provide the basic comprehension support to keep the student from feeling overwhelmed.

Target Instruction to Students' Abilities

To encourage optimal progress with the use of any reading materials, teachers need to be aware of the complexity level of the text relative to a student's reading level. A text that is too difficult may serve to undermine a student's confidence and diminish learning. Frequent use of text that is too easy may foster poor work habits and unrealistic expectations that will undermine the later success of the best students.

When students confront new kinds of texts and texts containing new content, the introduction can be softened and made less intimidating by guiding the student to easier reading. On the other hand, students who are comfortable with a particular genre or format or the content of such texts can be challenged with more difficult reading levels, which will reduce boredom and promote the greatest rate of development of vocabulary and comprehension skills.

Help Students Set Appropriate Learning Goals

Students' Lexile measures can be used to identify reading materials that students are likely to comprehend with the desired level of accuracy. Students can set goals of improving their reading comprehension and plan clear strategies for reaching those

goals using literature from the appropriate Lexile ranges. Progress tests throughout the year can help to monitor students' progress toward their goals.

Monitor Reading Program Goals

As a student's Lexile measure increases, the set of reading materials he can likely comprehend well changes. Schools can use the Lexile Framework for program review purposes. Schools can use student-level and school-level Lexile information to monitor and evaluate interventions designed to improve reading skills.

Measurable goals can be clearly stated in terms of Lexile measures. Examples of measurable goals and clearly related strategies for reading intervention programs might include.

Goal: At least half of the students will improve reading comprehension abilities by 100L after one year of use of an intervention.

Goal: Students' attitudes about reading will improve after reading 10 books at their targeted Lexile range.

These examples of goals emphasize the fact that the Lexile Framework is not an intervention, but a tool to help educators plan instruction and measure the success of the reading program.

Communicate with Parents Meaningfully to Include Them in the Educational Process

Teachers can make statements to parents such as, "Your child should be ready to read with adequate comprehension these kinds of materials which are at the next grade level." Or, "Your child will need to increase his/her Lexile measure by 400L–500L in the next few years to be prepared for college reading demands. Here is a list of appropriate titles your child can choose from for reading this summer."

Improve Students' Reading Fluency

Fluency is highly correlated to comprehension (Fuchs et al. 2001; Rasinski 2009). Educational researchers have found that students who spend a minimum of three hours a week reading at their own level for their own purposes develop reading fluency that leads to improved mastery. Not surprisingly, researchers have found that students who read age-appropriate materials with a high level of comprehension also learn to enjoy reading.

College, Career Readiness, and Text Complexity

There is increasing recognition of the importance of bridging the gap that exists between K-12 and higher education and other postsecondary endeavors. In the United States, many state and policy leaders have formed task forces and policy committees such as P-20 councils. In the *Journal of Advanced Academics* (2008), Williamson investigated the gap between U.S. high school textbooks and various reading materials across several U.S. postsecondary domains. The resources

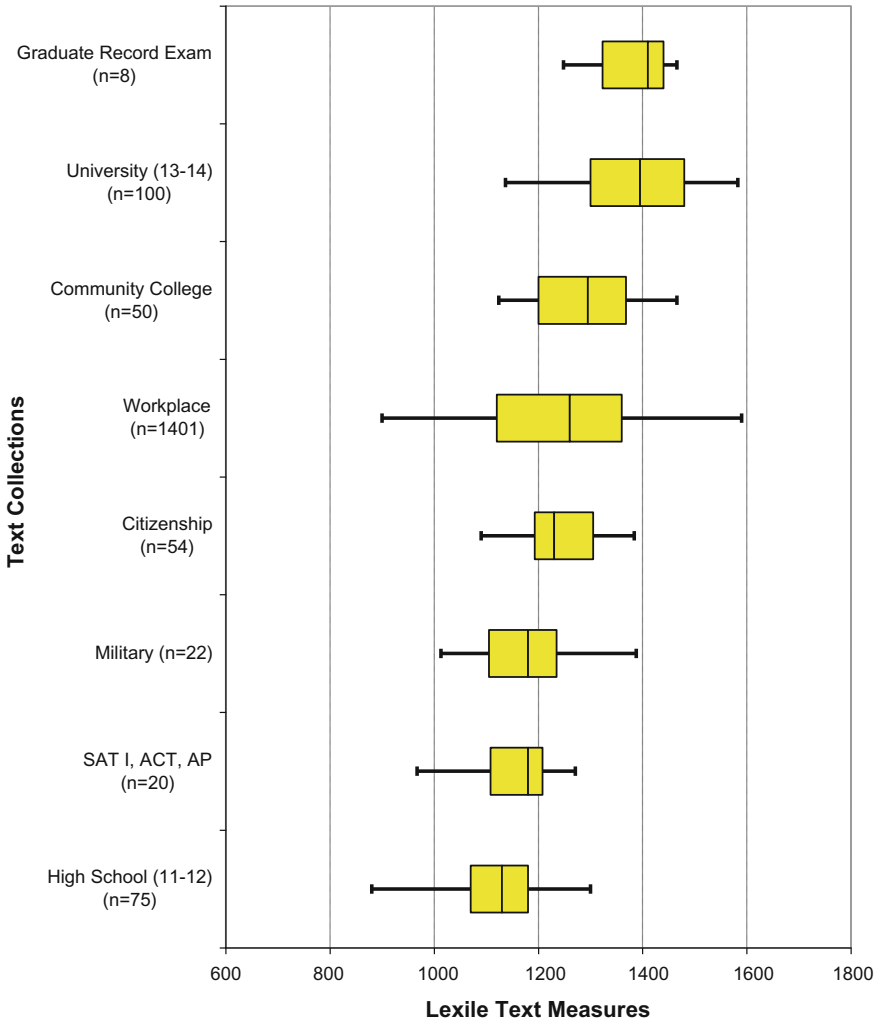


Fig. 2 A continuum of text difficulty for the transition from high school to postsecondary experiences (box plot percentiles. 5th, 25th, 50th, 75th, and 95th)

Williamson used were organized into four domains that correspond to the three major postsecondary endeavors that students can choose—further education, the workplace or the military, and, the broad area of citizenship, which cuts across all postsecondary endeavors. Williamson discovered a substantial increase in reading expectations and text complexity from high school to postsecondary domains—"a gap large enough to help account for high remediation rates and disheartening graduation statistics" (Smith 2011). Figure 2 illustrates this continuum of text difficulty.

Expanding on Williamson's work, Stenner et al. (2012) aggregated readability information across the various postsecondary options available to a high school graduate to arrive at a standard of reading needed by individuals to be considered "college and career ready." In their study, they included additional citizenship materials beyond those examined by Williamson (e.g., national and international newspapers and other adult reading materials such as Wikipedia articles). Using a weighted mean of the medians for each of the postsecondary options (education, military, work place, and citizenship), a measure of 1300L was defined as the general reading demand for postsecondary options and could be used to judge a student's "college and career readiness."

Subsequently, studies in Texas, Georgia, and Tennessee were conducted to examine the reading demands in various postsecondary options—technical college, community college, and 4-year university programs (MetaMetrics 2008). In terms of mean text demand, the results across the three states produced similar estimates of the reading ability needed in higher education institutions: Texas, 1230L; Georgia, 1220L; and Tennessee, 1260L. When these results are incorporated with the reading demands of other postsecondary endeavors (military, citizenship, workplace, and adult reading materials [national and international newspapers] and Wikipedia articles) used by Stenner et al. (2010), the college and career readiness standard for reading is 1293L. These results are based on more than 105,000,000 words from approximately 3100 sources from the adult text space.

Between 2004 and 2008, MetaMetrics (Williamson et al. 2012) collected and measured textbooks across the K-12 educational continuum. The box-and-whisker plot in Fig. 3 shows the Lexile measures (*y*-axis) across grades as defined in the U.S For each grade, the box refers to the interquartile range. The line within the box indicates the median. The end of each whisker shows the 5th and 95th percentile text complexity measures in the Lexile metric for each grade. This information can provide a basis for defining at what level students need to be able to read to be ready for various postsecondary endeavors in the United States such as further education beyond high school and entering the work force.

This continuum can be "stretched" to describe the reading demands expected of students in Grades 1–12 who are "on track" for college and career (Sanford-Moore and Williamson 2012). The quantitative aspect of defining text complexity consists of a stair-step progression of increasingly difficult text by grade levels (Common Core State Standards for English Language Arts, Appendix A, NGA Center and CCSSO 2010, p. 8).

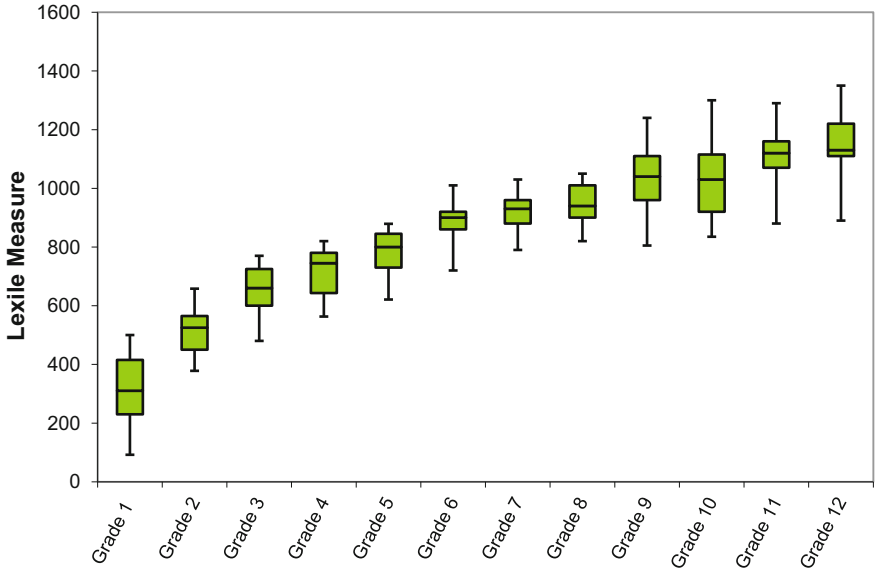


Fig. 3 Text complexity distributions, in Lexile units, by grade (whiskers represent 5th and 95th percentiles)

The question for educators becomes how to ensure that a student is ready for college and career reading demands. By making decisions about appropriate reading materials that progressively expose students to increasing challenge, educators can better prepare their students for their next levels of reading.

Conclusion

Just as variables other than temperature affect comfort, variables other than semantic and syntactic complexity affect reading comprehension. A student's personal interests and background knowledge are known to affect comprehension. However, although temperature alone does not fully identify the comfort level of an environment, we do not dismiss the importance of the information communicated by temperature. Similarly, the information communicated by the Lexile Framework is valuable, even though other information also enhances instructional decisions. In fact, the meaningful communication that is possible when test results are linked to instruction provides the opportunity for parents and students to give input regarding interests and background knowledge.

References

Bormuth, J. R. (1966). Readability: New approach. *Reading Research Quarterly, 7*, 79–132.

Carroll, J. B., Davies, P., & Richman, B. (1971). *Word frequency book*. Boston: Houghton Mifflin.

Carver, R. P. (1974). Measuring the primary effect of reading. Reading storage technique, understanding judgments and cloze. *Journal of Reading Behavior, 6*, 249–274.

Chall, J. S. (1988). The beginning years. In B. L. Zakaluk & S. J. Samuels (Eds.), *Readability: Its past, present, and future*. Newark, DE: International Reading Association.

Crain, S. & Shankweiler, D. (1988). Syntactic complexity and reading acquisition. In A. Davidson and G.M. Green (Eds.), *Linguistic complexity and text comprehension. Readability issues reconsidered*. Hillsdale, NJ. Erlbaum Associates.

Davidson, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable text: A case study from adaptations. *Reading Research Quarterly, 17*, 187–209.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test-revised: Forms L and M*. Circle Pines, MN: American Guidance Service.

Fuchs, L. S., Fuchs, D., Hops, M. K., & Jenkins, J. R. (2001). Oral reading as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–245.

Grolier, Inc. (1986). *The electronic encyclopedia*. Danbury, CT: Author.

Klare, G. R. (1963). *The measurement of readability*. Ames, IA: lowa State University Press.

Liberman, I. Y., Mann, V. A., Shankweiler, D., & Westelman, M. (1982). Children's memory for recurring linguistic and non-linguistic material in relation to reading ability. *Cortex, 18*, 367–375.

MetaMetrics, Inc. (2008). *Text measurement and analysis: MetaMetrics technical report update for the texas higher education coordinating board*. Durham, NC: Author.

Miller, G. A., & Gildea, P. M. (1987). How children learn words. *Scientific American, 257*, 94–99.

National Governors Association Center for Best Practices (NGA Center) & the Council of Chief State School Officers (CCSSO). (2010). *Common core state standards for English language arts and literacy in History/Social studies, Science and technical subjects: Appendix A*. Retrieved from http://www.corestandards.org/assets/Appendix_A.pdf.

Rasinksi, T.V. (2009). *Essential readings on fluency*. International Reading Association, Newark, DE.

Sanford-Moore, E., & Williamson, G. L. (2012). *Bending the text complexity curve to close the gap (MetaMetrics research brief)*. Durham, NC: MetaMetrics Inc.

Shankweiler, D., & Crain, S. (1986). Language mechanisms and reading disorder: A modular approach. *Cognition, 14*, 139–168.

Smith, M. (2011, March 30). *Bending the reading growth trajectory: Instructional strategies to promote reading skills and close the readiness gap (MetaMetrics policy brief)*. Durham, NC: MetaMetrics, Inc.

Stenner, A. J. (1990). Objectivity: Specific and general. *Rasch Measurement Transactions, 4*, 111.

Stenner, A. J., Koons, H., & Swartz, C. W. (2010, unpublished manuscript). *Text complexity and developing expertise in reading*. Durham, NC: MetaMetrics, Inc.

Stenner, A. J., Sanford-Moore, E., & Williamson, G. L. (2012). *The Lexile® framework for reading quantifies the reading ability needed for College & career readiness (MetaMetrics research brief)*. Durham, NC: MetaMetrics, Inc.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement, 20*(4), 305–315.

Stenner, A. J., Smith, D. R., Horiban, I., & Smith, M. (1987). *Fit of the Lexile theory to item difficulties on fourteen standardized reading comprehension tests*. Durham, NC: MetaMetrics Inc.

Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal of Advanced Academics, 19*(4), 602–632.

Williamson, G. L., Koons, H., Sandvik, T., & Sanford-Moore, E. (2012). *The text complexity continuum in grades 1–12 (MetaMetrics research brief)*. Durham, NC: MetaMetrics Inc.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.